Modelos de evaluación genómica con metafundadores y efectos de dominancia

Tesis para optar al título de Doctor de la Universidad de Buenos Aires
Área Ciencias Agropecuarias

Carolina Andrea García Baccino

Ing. Agr. – FAUBA - 2012

MS – FAUBA - 2017

Lugar de trabajo: Cátedra de Mejoramiento Genético Animal. Facultad de Agronomía. Universidad de Buenos Aires

Escuela para Graduados Ing. Agr. Alberto Soriano
Facultad de Agronomía – Universidad de Buenos Aires

COMITÉ CONSEJERO

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc (Montana State University, Estados Unidos de América)
MSc (University of Illinois, Estados Unidos de América)
PhD (University of Illinois, Estados Unidos de América)

Consejera de Estudios
Zulma Gladis Vitezica
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Universidad de Buenos Aires, Argentina)
Docteur (Institut National Agronomique Paris-Grignon, Francia)

JURADO DE TESIS

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc (Montana State University, Estados Unidos de América)
MSc (University of Illinois, Estados Unidos de América)
PhD (University of Illinois, Estados Unidos de América)

JURADO
Guillermo Giovambattista
Lic. Biol. (Universidad Nacional de La Plata, Argentina)
Doctor (Universidad Nacional de La Plata, Argentina)

JURADO
Daniel Omar Maizon
Méd. Vet. (Universidad de Buenos Aires)
MSc. (Universidad de Buenos Aires, Argentina)
PhD (Cornell University, Estados Unidos de América)

JURADO
Fernando Sebastián Baldi Rey
Ing. Agr. (Universidad de la Republica, Uruguay)
MSc. (Universidade Estadual Paulista, Brasil)
Doutor (Universidade Estadual Paulista, Brasil)

Fecha de defensa de la tesis: 11 de marzo de 2019

Nunca en mi vida había utilizado una herramienta, más con el tiempo, con trabajo, empeño e ingenio descubrí que no había nada que no pudiera construir, en especial, si tenía herramientas.

(Daniel Defoe**, "Robinson Crusoe")**

A mis padres por enseñarme el valor del trabajo y del esfuerzo, por brindarme los instrumentos para cimentar mi camino y la libertad y confianza para usar mis propias herramientas.

A mis abuelos y Matías por enseñarme el valor de dos herramientas que permiten construir una infinidad de cosas: el esfuerzo y la perseverancia.

AGRADECIMIENTOS

Como primera medida me gustaría agradecer a los miembros de mi comité consejero: Dr. Rodolfo Juan Carlos Cantet (Fito) y Dra. Zulma Vitezica (Zule). Fito, gracias por guiarme a lo largo de todos estos años, por sus consejos y enseñanzas a nivel profesional y personal, por permitirme formar parte de un grupo tan lindo como es el de MGA y por enseñarme el valor de trabajar de manera independiente con un espíritu crítico. Zule, muchas gracias por todo el apoyo, por guiarme todos estos años y por mostrarme a través del ejemplo el valor del esfuerzo y la perseverancia para alcanzar las metas. Gracias por todos los consejos y por la gran disposición y calidez con la que me recibieron, junto a Andrés, en Toulouse en dos oportunidades a lo largo del desarrollo de esta tesis. Fueron experiencias únicas e inolvidables de gran crecimiento personal y profesional.

A los miembros del jurado, Dr. Fernando Baldi, PhD. Daniel Maizon y Dr. Guillermo Giovambattista. Gracias por sus valiosos aportes a la versión final de esta tesis así como también a la discusión de los resultados obtenidos.

Al Dr. Sebastián Munilla Leguizamón. Gracias Sebas por acompañarme y apoyarme todos estos años. Por ayudarme a ver más allá de los obstáculos y seguir hacia adelante siempre.

Al Dr. Andrés Legarra. Gracias por darme la oportunidad de trabajar en este proyecto, pero sobre todo por la motivación, confianza y libertad. Gracias por todo el apoyo.

A la Facultad de Agronomía de la Universidad de Buenos Aires y a la cátedra de Mejoramiento Genético Animal por el apoyo institucional.

A CONICET por darme la posibilidad, a través de una beca, de iniciar y culminar esta tesis.

A INRA Toulouse e INP Toulouse por financiar mi estancia en Francia para terminar este trabajo. Al Ministerio de Educación de la República Argentina y al Ministerio de Asuntos Exteriores y Desarrollo Internacional de la República Francesa por otorgarme la Beca Saint-Exupéry que me permitió viajar a Francia para trabajar en la tesis.

A la American Angus Association por ceder el uso de sus bases de datos.

A Daniela Lourenco por sus valiosos comentarios y sugerencias.

A Marito por esos ratos de mates y risas por las tardes.

A Susana y Amelia, por todo su cariño y alegría.

A todo el personal de la Escuela para graduados entre los que me gustaría destacar a María Del Carmen Fabrizio por apoyarme en todo momento y por su dedicación, trabajo incansable y pasión por enseñar.

A mis compañeros y amigos de Biometría: María Isabel, Camilo y Pablo por acompañarme en todo el proceso de cursada y aprendizaje.

A mis compañeros y amigos de la Cátedra de Mejoramiento Genético Animal: Sebas, Nati, Andresito, Juan, Joselito, Yeni, Mati S., Dani, Mati B., Belcy, Majo, Esteban, Roberto y Martín. Gracias por todos sus consejos, apoyo y alegría cotidiana.

A mis compañeros y amigos de Francia: Patricia, Tomás, Fernando, David, Tara, Merina, Eli, Tao, Hung, Silvia, Alba y Oscar por la calidez con la que me recibieron y por la alegría cotidiana en esos días lejos de casa.

A Valeria Schindler, Laura Pruzzo, Mónica Santos Cristal por todas las enseñanzas que me transmitieron al inicio de mi camino en la cátedra. En especial para Anita Birchmeier por transmitirme muchas enseñanzas, pero por sobre todo su alegría y pasión al trabajar. Gracias por los consejos y apoyo a lo largo de todos estos años.

A mis amigos por acompañarme, aconsejarme y entenderme todos estos años.

A mis padres, Claudia y Daniel, por enseñarme desde el ejemplo a trabajar por lo que uno quiere y a esforzarse día a día hasta alcanzarlo. Gracias por su apoyo incondicional, por sus consejos y por darme todas las herramientas para construir mi camino.

A mis hermanos por acompañarme siempre con su alegría y consejos.

A Matías, por estar siempre, entenderme y acompañarme. Por ser mi apoyo en momentos decisivos e impulsarme a seguir.

A mis abuelos por enseñarme desde el ejemplo que por más obstáculos que surjan en el camino siempre hay un modo de seguir adelante. En especial a Coco quien sin querer me enseñó a cuestionar todo y a trabajar incansablemente hasta alcanzar mis objetivos.

A todos muchas gracias.

Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.

Carolina Andrea García Baccino

Publicaciones derivadas de la tesis

- **García-Baccino**, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G. y Cantet, R. J. C. 2017. Metafounders are related to F_{st} fixation indices and reduce bias in single step genomic evaluations. Genet., Sel., Evol.,. 49: 34.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

CUADRO página

ÍNDICE DE FIGURAS

FIGURA página

ABREVIATURAS

La mayoría de las abreviaturas empleadas provienen de los términos en inglés.

BLUE	estimador lineal insesgado de mínima varianza
BLUP	predictor lineal insesgado de mínima varianza
DS	desvío estándar
EBV	valor de cría predicho
EHW	equilibrio Hardy-Weinberg
F	coeficiente de consanguinidad
GBLUP	metodología BLUP de predicción del valor de cría genómico
GLS	mínimos cuadrados generalizados
IBD	identidad por descendencia
IBS	identidad por estado
LD	desequilibrio gamético o de ligamiento
LE	equilibrio gamético
MAF	frecuencia del alelo menos frecuente
MF	metafundador/metafundadores
ML	máxima verosimilitud
MME	ecuaciones de modelo mixto
MSE	error cuadrático medio
N_e	tamaño efectivo
PBLUP	BLUP basado en la genealogía
PEC	(co)varianza del error de predicción

QTL	locus que influye sobre un carácter cuantitativo
REML	Máxima verosimilitud restringida
SG	selección genómica
SNP	polimorfismo de un único nucleótido, marcador molecular
ssGBLUP	GBLUP en un solo paso
ssGBLUP_F	ssGBLUP considerando la consanguinidad en el cálculo de $\boldsymbol{A}^{-1}$
ssGBLUP_M	ssGBLUP incorporando un metafundador
TBV	valor de cría verdadero
UPG	grupos de padres desconocidos
WGR	modelo de regresión que utiliza todos los marcadores del genoma

Título: Modelos de evaluación genómica con metafundadores y efectos de dominancia

RESUMEN

Uno de los desafíos en el marco del GBLUP en un solo paso (ssGBLUP) es lograr la compatibilidad entre las matrices ***A*** y ***G*** al momento de calcular la matriz de estructura de covarianzas ***H***. Una posible solución es el empleo de metafundadores (MF), *pseudo* individuos que permiten cuantificar las relaciones de parentesco ancestrales de una o varias poblaciones. En esta tesis se generaron contribuciones teóricas y metodológicas con relación a la estimación de los parámetros centrales de un modelo con MF para simplificar su implementación. Por un lado, se mostró la relación teórica entre la relación ancestral (γ) con las covarianzas de las frecuencias alélicas de la población base y el índice de fijación, F_{st}. Además, se propusieron y evaluaron por simulación métodos que permiten emplear la información genómica de animales relacionados para estimar γ. Las mejores estimaciones (más exactas e insesgadas) se obtuvieron empleando Máxima Verosimilitud (ML) y Mínimos Cuadrados Generalizados (GLS). Adicionalmente, se presentó un modo sencillo para calcular el parámetro *s*, relacionado con la heterocigosidad de los marcadores, en función del número de SNP empleados en el análisis. Se evaluó también el impacto predictivo de incorporar un MF en ssGBLUP por simulación estocástica, así como también en términos de la estimación de los componentes de varianza. Como resultado se encontró que las predicciones empleando MF presentaron similar exactitud y menor sesgo que las obtenidas con ssGBLUP tradicional. Además, el modelo con MF permitió obtener estimaciones de los parámetros genéticos más precisas. Finalmente, se evaluó la incorporación de los efectos de dominancia en un modelo de predicción para caracteres de crecimiento con información genómica en una población real de bovinos de carne. Los resultados sugieren que la proporción de la varianza genética explicada por dominancia es pequeña, y que el ajuste del modelo no mejora al considerarla. Se estimaron también valores de depresión consanguínea para los caracteres evaluados.

Palabras claves: selección genómica, metafundadores, relaciones ancestrales, estimación de parámetros, predicción, dominancia, componentes de (co)varianza.

Title: Genomic evaluation models with metafounders and dominance effects.

ABSTRACT

A challenge in Single Step GBLUP (ssGBLUP) is to achieve compatibility between matrices $\boldsymbol{A}$ and $\boldsymbol{G}$ when computing the covariance matrix $\boldsymbol{H}$. One possible solution is the use of metafundadores (MF), *pseudo*-individuals that describe relationships within and across pedigree base populations. In this thesis, theoretical and methodological contributions were generated for the estimation of the central parameters of a model including MF to simplify its implementation. We showed that ancestral relationship parameters (γ) are proportional to standardized covariances of base allelic frequencies across populations, such as F_{st} fixation indexes. These covariances of base allelic frequencies can be estimated from marker genotypes of related recent individuals and pedigree. Different methods to estimate γ were proposed and their performances were assessed by simulation. We observed that generalized least squares (GLS) or maximum likelihood (ML) gave accurate and unbiased estimates of the ancestral relationship parameter (γ). Additionally, a simple way of estimating the parameter *s*, related to the heterozygosity of the markers, based on the number of SNPs used in the analysis was shown. The quality of genomic predictions and variance component estimation using MF was also tested. Inclusion of MF relationships reduces bias of genomic predictions with no loss in accuracy and produces consistent estimates of heritability. Finally, we estimated additive and dominance variance components in a real beef cattle population for growth traits taking into account inbreeding depression and genomic information. Results show that the proportion of the genetic variance explained by dominance is small. Additionally, for all traits the model with only additive effects fitted the data better than the one including non-additive effects. Inbreeding depression was also estimated for the growth traits.

Key words: genomic selection, metafounders, ancestral relationships, parameter estimation, prediction, dominance, (co)variance components.

Capítulo 1

Introducción general

Introducción general

1.1. INTRODUCCIÓN

A medida que la sanidad, la nutrición y el manejo dejan de ser limitantes a la producción animal, la selección de reproductores es el principal elemento para aumentar la producción o su eficiencia. El progreso genético de una población animal es fundamentalmente generado por la selección sostenida a lo largo de los años, buscando emplear como progenitores aquellos animales de mérito genético superior. En las evaluaciones genéticas tradicionales, la predicción de los valores de cría se realiza mediante un sólido fundamento estadístico como es la teoría de Predicción Lineal Insesgada de Mínima Varianza (BLUP, Henderson, 1984); teoría que combina apropiadamente los registros fenotípicos y la genealogía para producir las predicciones. En años recientes los avances en el área de la genética molecular permitieron disponer de grandes volúmenes de información genómica, resultado del advenimiento de las técnicas de identificación de marcadores SNP (del inglés "Single Nucleotide Polymorphysm", polimorfismos genéticos de un solo nucleótido) en alta densidad. Como consecuencia, surgió el desafío de incorporar dicha información a las evaluaciones genéticas de modo de contar con predicciones más precisas. Esto introdujo un cambio en la predicción del mérito genético, dando lugar al desarrollo de los modelos de "selección genómica" (SG, Meuwissen *et al.*, 2001), hecho que ha generado un cambio paradigmático en el área de la mejora genética animal.

Meuwissen *et al.* (2001) propusieron un modelo de regresión que utiliza todos los marcadores del genoma (WGR). Para un carácter de interés con herencia poligénica, los efectos del modelo son los marcadores responsables de captar la variabilidad genética asociada - por desequilibrio gamético - entre marcadores y genes. Una vez estimados los efectos, se los suma sobre la base del genotipo que posea cada individuo en cada SNP. El resultado es el valor de cría predicho "genómico" para cada individuo genotipado. Ahora bien, los modelos WGR sufren la limitación de que los efectos se definen sobre los SNP pero los marcadores no son genes. Consecuentemente, deben ser considerados como modelos de "error de medición" (de los Campos *et al.*, 2015), y no pueden utilizar toda la información de similitud genética entre individuos. Además, y dado que los SNP son marcadores bialélicos, el modelo suele presentar gran variabilidad en las estimaciones de los efectos de los marcadores individuales cuando la frecuencia del alelo menos común tiende a cero. A los efectos de remediar este problema predictivo, se propusieron diversos modelos alternativos de SG; por ejemplo, regularizar los efectos de los marcadores empleando diferentes distribuciones o mezclas (de los Campos *et al.*, 2009). Al asumir una distribución gaussiana de los efectos de los marcadores, es posible reformular el modelo y emplear las ecuaciones de modelos mixtos (MME) clásicas de C. R. Henderson (GBLUP). La única modificación necesaria para considerar la información molecular en este caso es reemplazar la matriz $\boldsymbol{A}$ de relaciones aditivas calculadas condicional a la genealogía por la

G, de relaciones genómicas. Esta última matriz puede calcularse considerando información genealógica y molecular (identidad por descendencia, IBD; Malecot, 1948) o únicamente molecular (identidad en estado - IBS).

Un problema complejo al momento de incorporar la información molecular en la predicción del valor de cría es que no todos los animales de una población se encuentran genotipados, sea por cuestiones logísticas o por costo. Desde el punto de vista de la predicción entonces existen dos tipos de individuos: los genotipados y aquellos sin información molecular. Para abordar esta realidad se propusieron, primero, evaluaciones genómicas en varias etapas (VanRaden, 2008; VanRaden *et al.,* 2009), que generaban sesgos e ineficiencias a la hora de emplear toda la información disponible. Posteriormente surgió la evaluación genómica en un único paso, comúnmente conocida como "*Single-step*" (ssGBLUP) introducida por Misztal *et al.* (2009), Legarra *et al.* (2009) e, independientemente, por Christensen y Lund (2010). Este procedimiento representa una alternativa práctica para combinar la información genómica (matriz ***G***) y genealógica (matriz ***A***), dentro de una matriz de relaciones ***H*** (Legarra *et al.*, 2014). Las relaciones genómicas pueden proyectarse a través del pedigrí para animales que no cuentan con información genómica (Legarra *et al.*, 2009 y Christensen y Lund, 2010). A tal efecto, se requiere que las relaciones genómicas y de pedigrí se refieran a la misma "base" (Legarra *et al.*, 2015). La "población base" o, simplemente la "base", es aquel conjunto de individuos de la población ancestral fundadora del pedigrí y comúnmente se los asume no relacionados entre sí. Esto representa un problema porque la base es difícil de definir y no suelen coincidir para ambos tipos de relaciones, lo que da lugar a incompatibilidades entre las matrices ***A*** y ***G***.

Mientras que la referencia para las relaciones de pedigrí está dada por fundadores de la genealogía, la de las relaciones genómicas suele ser el conjunto de animales genotipados (Powell *et al.*, 2010; Vitezica *et al.*, 2011). Las relaciones genómicas de la población actual cambian a medida que se ingresan nuevos animales genotipados y están pobremente definidas si las poblaciones poseen estructura ya que no logran captarla en su totalidad. Por su parte, también es complicado definir la base para relaciones de pedigrí dado que los pedigríes son incompletos y pueden remontarse a más de una población ancestral (Legarra *et al.*, 2015). Powell *et al.* (2010) mostraron que conceptualmente es posible referir las relaciones genómicas a la escala de las de pedigrí y viceversa. Nociones similares fueron introducidas por VanRaden (2008) y Vitezica *et al.* (2011). Dichos autores propusieron modificar las relaciones genómicas para referirlas a las genealógicas empleando un supuesto implícito: la población genotipada no posee estructura genealógica (Christensen, 2012).

A su vez, las relaciones de pedigrí poseen varios problemas (Legarra *et al.*, 2015). Las genealogías suelen ser incompletas y se remontan a una o varias poblaciones base (líneas o razas). Tradicionalmente se asume que los individuos base de una población no se encuentran relacionados entre sí dado que provienen de una población ancestral de gran

tamaño. Sin embargo, la disponibilidad de información molecular permitió demostrar que este supuesto no se cumple para la mayoría de las poblaciones reales (Legarra *et al.*, 2015). Por otro lado, se observan casos en que las genealogías se encuentran incompletas diferencialmente según la categoría o sexo de los animales. Por ejemplo, en bovinos de carne suele ocurrir que los padres de los machos son conocidos en una mayor proporción que para el caso de las hembras. Además, asumir que en presencia de selección, todos los padres desconocidos pertenecen a la misma población base y poseen el mismo nivel genético es incorrecto puesto que los animales más jóvenes son selectos y, esperablemente mejores que aquellos en la base. Si el modelo de predicción no toma en cuenta el impacto de la selección e información faltante sobre la distribución del carácter evaluado, pueden aparecer sesgos predictivos (Misztal *et al.*, 2013). El empleo de grupos de padres desconocidos o grupos genéticos surgió como una alternativa para lidiar con esta problemática: los padres desconocidos suelen asignarse a diferentes poblaciones base tomando como criterio de clasificación el año de nacimiento o el país de origen del animal, entre otros clasificadores. Se asume que los diferentes grupos poseen valores promedio a priori diferentes, y se estiman como efectos fijos dentro del modelo (Thompson 1979 y Quaas 1988). Kennedy (1991) señaló que dichos grupos se asumen incorrectamente no relacionados entre sí.

Christensen (2012) propuso tomar una referencia arbitraria para las relaciones genómicas y emplear una población ideal con frecuencias alélicas de 0,5 para los marcadores. Además, propuso referir las relaciones de pedigrí a dicha población base. De esto modo mostró que los fundadores de la población base deberían estar relacionados, y que dicha relación puede interpretarse como un exceso de homocigosidad IBD. Legarra *et al.* (2015) trabajaron sobre esta propuesta y presentaron una teoría para considerar las relaciones entre y dentro poblaciones bases o fundadores. A tal fin introdujeron la noción del "metafundador" (MF) que permite condensar la información de parentescos en la población ancestral a través de la estimación del parámetro γ. Su utilización permite modificar las matrices de parentescos calculadas por genealogía ($\boldsymbol{A}$) con el objetivo de ajustarlas a las genómicas ($\boldsymbol{G}$), permitiendo compatibilizarlas. La teoría de Legarra *et al.* (2015) surgió como una extensión de los trabajos de Jacquard (1969, 1974), VanRaden (1992), Aguilar y Misztal (2008), VanRaden *et al.* (2011), Colleau y Sargolzaei (2011) y Christensen (2012). La teoría proveyó los fundamentos para generalizar el empleo de los grupos de padres desconocidos y los resultados de Christensen (2012). Los conceptos desarrollados por Legarra *et al.* (2015) son de interés en dos aspectos: 1) al combinar relaciones genómicas y de pedigrí, tal como ocurre en ssGBLUP y 2) al considerar casos con varias poblaciones base. Legarra *et al.* (2015) presentaron el marco teórico y metodológico sin evaluar su desempeño en términos predictivos y de la estimación de los parámetros centrales del modelo. Es en estos puntos donde esta tesis pretende realizar sus aportes contribuyendo a evaluar la performance del método en términos prácticos y generando propuestas metodológicas para la estimación de los parámetros centrales del modelo.

Adicionalmente, la creciente disponibilidad de información molecular contribuyó también a renovar el interés por considerar los efectos no aditivos dentro de los modelos de selección genómica. Tradicionalmente, la tendencia fue a ignorarlos y dicha propensión aún se observa en la actualidad con los modelos genómicos. Los motivos que llevaron a ignorarlos son, principalmente, de índole operativo y práctico, tal como los presenta Varona *et al.* (2018). Entre ellos se destaca la falta de pedigríes informativos, en especial, grandes familias de hermanos enteros, la complejidad de los cálculos y demanda computacional asociada con los mismos. Además, el hecho de que la varianza aditiva capture efectos no aditivos (Hill, 2010), contribuye a no considerar directamente la dominancia y epistasis dentro del modelo, pese a que la primera puede explicar una parte significativa de la varianza genética (Varona *et al.*, 2018). En este sentido, el hecho de incorporar los efectos no aditivos en el modelo puede acarrear ciertas ventajas como aumentar la exactitud de predicción de los valores de cría y la respuesta a la selección (Toro y Varona, 2010; Aliloo *et al.*, 2016; Duenk *et al.*, 2017). Además puede resultar útil a la hora de definir apareamientos que permitan maximizar la performance productiva de los animales de la próxima generación considerando tanto el valor de cría como el de dominancia (Maki-Tanila, 2007; Toro y Varona, 2010; Aliloo *et al.*, 2017).

Dadas estas ventajas y, sumado al hecho de contar con crecientes volúmenes de información molecular adicional, en los últimos años se renovó el interés por considerar los efectos no aditivos dentro de los modelos de selección genómica. Toro y Varona (2010), Su *et al.* (2012) y Vitezica *et al.* (2013) abordaron dicho desafío y propusieron diferentes parametrizaciones para considerar los efectos no aditivos en los modelos incorporando el empleo de la información genómica. Dichos modelos fueron implementados en diferentes poblaciones de animales domésticos para diversos caracteres de interés económico según cada sistema productivo. Los resultados, en muchos casos fueron ambiguos y variables. Para el caso puntual de los bovinos de carne, se ha generado poca investigación reciente referida al tema. Esto puede atribuirse principalmente a la falta de grandes bases de datos con un gran número de animales con información genómica y fenotípica, tal como discute Varona *et al.* (2018). Recientemente, Bolormaa *et al.* (2015) evaluaron la factibilidad de incluir efectos no aditivos en el modelo de evaluación genómica en bovinos de carne. A tal fin consideraron varios caracteres pero sin centrar su atención en aquellos de alto impacto en términos productivos y económicos para este tipo de sistemas de producción como son los de crecimiento y calidad de res. Además, llevaron a cabo su análisis empleando datos de animales de diferentes razas y cruzas. Existen en la literatura estimaciones previas para dichos caracteres basadas en datos genealógicos y fenotípicos únicamente, con resultados variables y, en muchos casos, ambiguos o poco informativos.

1.2. OBJETIVOS GENERALES Y NATURALEZA DE LA TESIS

Sobre la base de lo anteriormente expuesto, la mayor motivación para el desarrollo de esta tesis surge de los últimos avances en genética poblacional y molecular que dieron lugar a un cambio paradigmático en la predicción del mérito genético. El desafío que implica incorporar la información molecular en las evaluaciones genéticas ha generado un gran interés y, aún se continúa investigando activamente en el tema con el objetivo de lograr un uso eficiente de dichos datos asegurando una implementación viable de las metodologías propuestas. Por lo tanto, en esta tesis se pretende abordar algunas de las problemáticas actuales de las evaluaciones genómicas y contribuir con propuestas de utilidad práctica y teórica.

Los objetivos generales de esta tesis son: 1) evaluar el empleo de MF en selección genómica, comparando su desempeño en términos predictivos con metodologías ampliamente utilizadas en la actualidad; 2) Contribuir al marco teórico del modelo con MF en términos de la definición de los parámetros centrales y el modo de calcularlos; 3) Proponer metodologías para la estimación de los parámetros centrales del modelo con MF con el objetivo de simplificar su implementación en evaluaciones genómicas reales; 4) Explorar la posibilidad de incluir efectos de dominancia en evaluaciones genómicas de bovinos de carne para caracteres de crecimiento. De estos objetivos generales, se desprenden varios objetivos específicos, que serán abordados en diferentes capítulos.

Esta tesis consta de siete capítulos, entre los que se incluye la presente introducción general. Luego, en el capítulo 2 se presenta el marco teórico más relevante de los MF y se describen dos contribuciones teóricas de relevancia para la implementación del método en selección genómica. El mismo será de referencia continua para los dos capítulos posteriores (3 y 4), dado que dependen de los resultados teóricos presentados en el segundo capítulo. En el capítulo 3 se proponen métodos alternativos a los existentes para la estimación de los parámetros centrales del modelo de selección genómica con MF. Además, se obtienen estimaciones de dichos parámetros con el objetivo de evaluar el desempeño de los diferentes métodos, sean los presentados inicialmente en la literatura (Legarra *et al.*, 2015), o aquellos propuestos en esta tesis. A tal efecto se emplea una base de datos simulados de una población bovina de leche bajo selección. Luego, en el capítulo 4 se evalúa la calidad de las predicciones genómicas y la estimación de componentes de varianza empleando MF en el modelo. Nuevamente, en este caso se emplea la base de datos simulada, mencionada previamente. Los resultados se comparan con aquellas predicciones generadas por métodos tradicionales y genómicos ampliamente utilizados en la actualidad. En el capítulo 5 se evalúa la posibilidad de incluir efectos de dominancia en evaluaciones genómicas de bovinos de carne para caracteres de crecimiento. A tal fin se emplearon datos reales de una población de bovinos de carne. Finalmente en los capítulos 6 y 7 se presenta la discusión general de la tesis y las conclusiones, respectivamente, así como también los aspectos pendientes para abarcar en investigaciones futuras.

Capítulo 2

Compatibilidad entre relaciones genómicas y genealógicas mediante el empleo de metafundadores

Compatibilidad entre relaciones genómicas y genealógicas mediante el empleo de metafundadores

2.1. INTRODUCCIÓN

Los Metafundadores (MF) (Legarra *et al.*, 2015) son individuos ficticios (*pseudo* individuos) que se adicionan al pedigrí de modo de explicar la ancestría en común entre individuos que, sobre la base de la información con que se cuenta comúnmente, no estarían emparentados. El concepto de MF intenta proveer un marco coherente para la teoría de la evaluación genómica. En especies domésticas de animales o plantas suelen llevarse a cabo actualmente evaluaciones genómicas en las que comúnmente sólo parte de los individuos se encuentran genotipados en alta densidad. A su vez, no todos los individuos cuentan con información fenotípica de manera que se puede contar con animales en cuatro situaciones diferentes: 1) con genotipo y fenotipo; 2) con genotipo, sin fenotipo; 3) sin genotipo, con fenotipo; 4) sin genotipo ni fenotipo. Con el objetivo de integrar todos los casos y lograr combinar las fuentes de información Legarra *et al.* (2009), Christensen y Lund (2010) y Fernando *et al.* (2014) propusieron una solución llamada Predicción lineal insesgada genómica en un solo paso o comúnmente conocida como "*Single-Step GBLUP"* (ssGBLUP del inglés *Single-Step genomic best linear unbiased prediction*). Dicha solución emplea la siguiente matriz de relaciones:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\left(\boldsymbol{G} - \boldsymbol{A}_{22}\right)\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G} \end{bmatrix} \qquad [2.1]$$

La inversa de la matriz $\boldsymbol{H}$ es

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \qquad [2.2]$$

donde $\boldsymbol{G}$ es la matriz de relaciones genómicas, $\boldsymbol{A}$ es la matriz de relaciones aditivas (Henderson, 1984), calculadas condicionales a la información de pedigrí y las matrices $\boldsymbol{A}_{11}$, $\boldsymbol{A}_{12}$, $\boldsymbol{A}_{21}$ y $\boldsymbol{A}_{22}$ son submatrices de $\boldsymbol{A}$ en las que el subíndice 1 refiere al grupo de animales no genotipados y el 2, a los que cuentan con información genómica.

Dado que comúnmente los animales genotipados no son una muestra representativa de la población analizada (los mismos tienden a ser en su mayoría jóvenes o animales selectos), se advirtió rápidamente la necesidad de un análisis apropiado de los datos que permitiera tomar en cuenta distintas medias para animales genotipados y no genotipados para el o los caracteres bajo análisis. Dichas medias pueden ser consideradas como

parámetros del modelo, ya sea como efectos fijos (Fernando *et al.*, 2014) o aleatorios (Vitezica *et al.*, 2011 y Christensen *et al.*, 2015). En el último caso, las variables aleatorias presentan covarianzas entre individuos, fenómeno al que informalmente se refiere con "compatibilidad" de las matrices de relaciones de pedigrí y genómicas. De hecho, la compatibilidad implica equidad en términos del promedio del valor de cría de la población base y de la varianza genética (Legarra, 2016) entre las distintas medidas de relación de parentesco. Numéricamente el problema se presenta como sigue. La fórmula para la matriz $\boldsymbol{H}$ y su inversa contienen los términos $\boldsymbol{G}-\boldsymbol{A}_{22}$ y $\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1}$ (asumiendo que $\boldsymbol{G}$ es de rango completo), respectivamente. Esto sugiere que sí $\boldsymbol{G}$ y $\boldsymbol{A}_{22}$ son muy distintas pueden presentarse sesgos.

Las relaciones genómicas son comúnmente calculadas empleando productos cruzados (VanRaden, 2008) o IBS corregido (Ritland, 1996). Ambos dependen críticamente de las frecuencias alélicas asumidas para la población base (Toro *et al.*, 2011). Las frecuencias alélicas de la población base usualmente son desconocidas. De todos modos, para la mayoría de los propósitos las mismas no son de interés *per se* y pueden ser tratadas como parámetros marginalizables. Christensen (2012) logró una integración algebraica de las frecuencias alélicas llegando a una estructura de covarianza muy simple con las frecuencias fijas en 0,5 (por ejemplo empleando –1, 0 y 1 para los códigos de los genotipos en el método de los productos cruzados) y un parámetro γ que describe las relaciones entre los fundadores del pedigrí. Es así que la matriz de relaciones aditivas de la población base puede calcularse como $\boldsymbol{A}_{base}^{\gamma}=\boldsymbol{I}\left(1-\frac{\gamma}{2}\right)+\boldsymbol{1}\boldsymbol{1}'\gamma$. Esta matriz señala un parentesco similar entre los fundadores que se encuentra dado por el parámetro γ. Existe un segundo parámetro en la marginalización de Christensen (2012), conocido como *s*. El mismo es una medida de la heterocigosidad en la población base. Consecuentemente, en lugar de inferir una gran cantidad (en el orden de las miles) de frecuencias alélicas base, sólo es necesario estimar dos simples parámetros: γ y *s* para computar la matriz de estructura de covarianzas propuesta por Christensen (2012). Ambos pueden estimarse maximizando la verosimilitud de los genotipos observados. Además, este enfoque considera el hecho de que la cantidad de información genealógica es variable y depende en gran medida de la disponibilidad histórica de registros.

Legarra *et al.* (2015) mostró la equivalencia entre el enfoque de Christensen (2012) y el concepto de los MF. Estos últimos pueden definirse como *pseudo* fundadores que encapsulan tres ideas: 1) diferentes medias para cada población base (Fernando *et al.*, 2014; Thompson, 1979 y Quaas, 1988), 2) aleatoriedad de dichas medias (Vitezica *et al.* 2011) y 3) propagación de esta aleatoriedad a la progenie (Christensen, 2012). Legarra *et al.* (2015) también generalizaron una sola relación entre fundadores (escalar γ) a varias relaciones entre fundadores en el pedigrí, es decir a diversas relaciones ancestrales (matriz **Γ**), y sugirieron métodos simples para estimarlas. Legarra *et al.* (2015) mostraron que la

construcción de $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ a partir de $\boldsymbol{\Gamma}$ y de la genealogía puede realizarse empleando las reglas del método tabular (Emik y Terril, 1949) para la construcción de relaciones y su inversión puede lograrse al invertir la matriz $\boldsymbol{\Gamma}$ y aplicando las reglas de Henderson (1976). Más adelante, en la sección 2.2.2.2. de este capítulo, se presenta un ejemplo de las matrices $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ y $\boldsymbol{\Gamma}$. El desempeño del modelo con MF propuesto por Legarra *et al.* (2015) no fue probado ya sea en términos de estimación de las relaciones ancestrales (γ) o para la predicción de valores de cría genómicos.

En este capítulo se presentan los fundamentos teóricos de los modelos de selección genómica incorporando MF propuestos por Legarra *et al.* (2015), quienes extendieron y generalizaron las ideas de Christensen (2012). Se introduce el marco teórico de los MF y el modo de incorporarlos en las evaluaciones genómicas en casos con una sola población base (un solo MF) o más de una (varios MF). Además, se presentan las propiedades de los MF, sus ventajas y el modo en que se relacionan a conceptos ampliamente utilizados dentro de la genética cuantitativa y de poblaciones. Es dentro de este último aspecto en el que se realizan las dos contribuciones teóricas originales de este capítulo: se demuestra que los parámetros de relaciones ancestrales son proporcionales a las covarianzas estandarizadas de las frecuencias alélicas base entre poblaciones y su relación con el índice de fijación F_{ST} (Wright, 1943). Estas covarianzas de las frecuencias alélicas base pueden estimarse empleado información de marcadores moleculares de generaciones recientes de las poblaciones, además de la información de genealogía. A tal efecto, se proponen en el capítulo 3 diferentes métodos de estimación de los parámetros que emplean información molecular de animales relacionados. Es importante destacar, que hasta el momento no se han propuesto metodologías que permitan inferir coeficientes F_{ST} a partir de información molecular con individuos emparentados. En consecuencia, esta constituye otra contribución teórica original de esta tesis. Los métodos propuestos fueron evaluados empleando datos simulados y su performance se comparó con aquellos previamente propuestos por Legarra *et al.* (2015). Los resultados permitieron determinar cuáles de ellos funcionan mejor con el objetivo de recomendar su uso en futuros trabajos o implementaciones de las evaluaciones genéticas empleando información genómica. De hecho, permitió definir el mejor método de estimación para los parámetros de interés para utilizar al momento de evaluar el impacto que genera - en términos predictivos - incorporar un MF en ssGBLUP. Los resultados de dicha prueba se presentan en el capítulo 4 y se contrastan con las predicciones generadas por otros tres métodos.

2.2. MARCO TEÓRICO Y METODOLOGÍA

En esta sección se presenta la base teórica de los MF propuesta por Legarra *et al.* (2015). A tal fin se resumen los aspectos de mayor relevancia presentados por los autores con el objetivo de comprender el marco teórico y posteriormente realizar contribuciones originales de carácter metodológico, teórico y de aplicación. Cabe destacar que la teoría de

los MF surgió como una extensión de los trabajos de Jacquard (1969, 1974), VanRaden (1992), Aguilar y Misztal (2008), VanRaden *et al.* (2011), Colleau y Sargolzaei (2011) y Christensen (2012).

2.2.1. Empleo de un metafundador en casos con una única población base

Antes de proceder a detallar la teoría propuesta por de Legarra *et al.* (2015), es necesario definir inequívocamente algunos conceptos referidos a las poblaciones que serán utilizados de aquí en adelante. A tal fin, adoptaremos y ampliaremos las definiciones presentadas por Legarra *et al.* (2015). Por un lado, denotaremos como "ancestral" a aquella población de la cual provienen las gametas que dieron origen a los individuos de la base. Otra posibilidad es considerar un conjunto ("*pool"*) de gametas que originaron a los individuos de la "base". La "población base" o, simplemente la "base", es aquel conjunto de individuos de la población ancestral fundadora del pedigrí. Ambas poblaciones se indican y diferencian en la Figura 2.1.

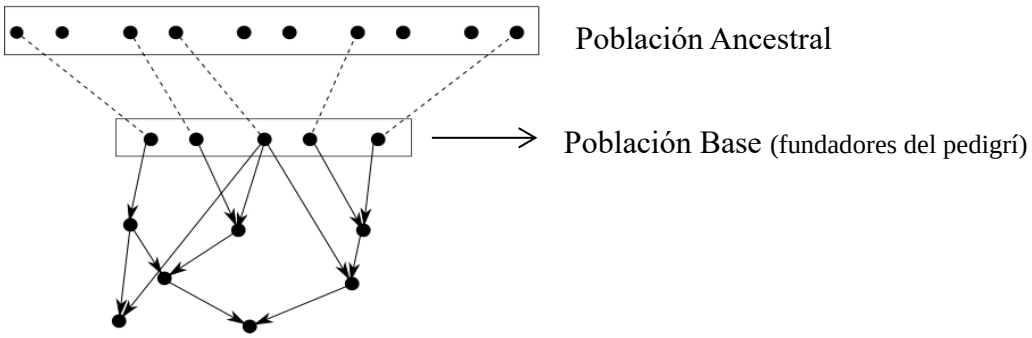


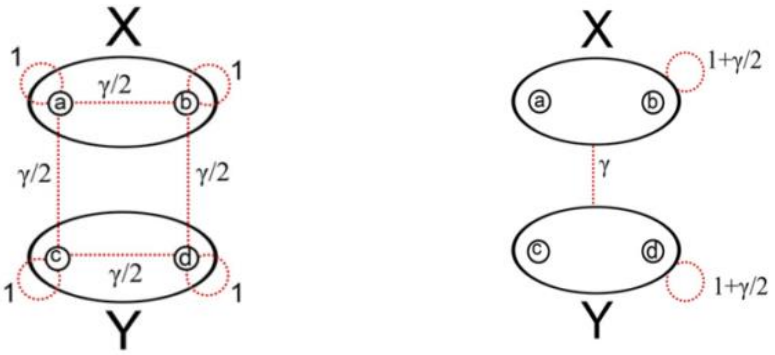
Figura 2.1. Esquema de la estructura de una población en la que se aprecia la población ancestral, de donde provienen las gametas originarias, la población base y el pedigrí. Modificada a partir de Legarra *et al.* (2015).

2.2.1.1. Relaciones entre los individuos de la población base

Tradicionalmente se asume que los individuos de la población base provienen de una población ancestral de gran tamaño con apareamientos aleatorios, tal que los individuos no se encuentren relacionados entre sí (i.e. las gametas originarias de la base provienen de individuos no emparentados). Ahora bien, tal como fuera mencionado anteriormente en el capítulo 1, existen evidencias empíricas que demuestran que dicho supuesto no se cumple en la mayoría de las poblaciones reales, dado que a partir del empleo de marcadores moleculares se suelen detectar relaciones entre individuos base. Entre los ejemplos pueden

mencionarse los trabajos de Kijas *et al.* (2009) en ovejas, Gibbs *et al.* (2009) y VanRaden (2011) en bovinos. En un primer abordaje del tema, y previo a la disponibilidad de información molecular, Jacquard (1969 y 1974) mostró que, en poblaciones de tamaño finito, tanto las relaciones de parentesco como la consanguinidad aumentan rápidamente a lo largo de las generaciones, lo que impacta directamente en el tamaño efectivo, tal como observaron en poblaciones vacunas Gibbs *et al.* (2009). Legarra *et al.* (2015) abordaron el enfoque de Jacquard (1969 y 1974) de un modo más sencillo tal como se describe a continuación.

Considerando un enfoque más amplio y tras el objetivo de contemplar la posibilidad que los individuos base se encuentren relacionados entre sí, es posible abordar el problema como un proceso de muestreo con reposición tal como lo presentan Legarra *et al.* (2015). Dichos autores plantean que los individuos de la población base (fundadores del pedigrí) son tomados al azar, con reposición, de una población ancestral finita con un tamaño efectivo (N_e) dado y 2 N_e gametas. Se asume que en dicha población ancestral las gametas son independientes, no relacionadas entre sí. Ahora bien, considérese en este escenario a dos gametas tomadas al azar con reposición de la población ancestral para formar la población base. La probabilidad de que ambas sean idénticas es $1/(2\,N_e)$. Nótese, además que en aquellos casos reales donde el N_e de la población ancestral es pequeño, mayor es la probabilidad de observar un evento en el cual la primera gameta muestreada es idéntica a la segunda. En consecuencia, la probabilidad de identidad por descendencia o IBD entre todos los pares de gametas es $\gamma/2 = 1/(2\,N_e)$ y se corresponde a la correlación entre gametas propuesta por Wright (1922). Posteriormente, Jacquard (1974) definió a α como $\alpha = \gamma/2$ y lo llamó "coeficiente de consanguinidad de una población" (Legarra *et al.* 2015). Las relaciones entre gametas se representan gráficamente en la Figura 2.2 (*i*), donde cada gameta se representa con las letras *a*, *b*, *c* y *d*, y los individuos se representan con las letras X e Y (nótese que cada individuo posee dos gametas).



(*i*) Nivel gamético **(*ii*)** Nivel de individuo

Figura 2.2. Esquema de las relaciones (representadas por líneas punteadas) entre dos individuos X e Y evaluadas a dos niveles: (*i*) gamético y (*ii*) de individuo, para el caso de una población base con individuos relacionados entre sí. El individuo X posee las gametas *a* y *b*, mientras que el Y, a *c* y *d*. En (*i*) las relaciones entre gametas es γ/2 y de las gametas consigo mismas es de 1. En (*ii*) las relaciones entre animales corresponden a γ y las del individuo consigo mismo es de 1+ γ/2. Tomada y modificada de Legarra *et al.* (2015).

Las relaciones entre individuos en la población base se representan en la Figura 2.2 (*ii*). Las gametas que posee cada uno fueron tomadas de un pool en el que la probabilidad de IBD es $\gamma/2$ entre gametas distintas y uno consigo misma, tal como se observa en la Figura 2.2 (*i*). En consecuencia, el coeficiente de coancestría entre X e Y (r_{XY}) puede calcularse como

$$r_{XY} = \frac{P(a \equiv c) + P(a \equiv d) + P(b \equiv c) + P(b \equiv d)}{4} = \frac{4(\gamma/2)}{4} = \gamma/2 \qquad [2.3]$$

Por su parte, la relación aditiva entre X e Y corresponde a dos veces la coancestría entre ellos, es decir, $2r_{XY} = \gamma$ (Figura 2.2 (*ii*)). Ahora bien, al tomar uno de los individuos, por ejemplo X, se observa que la relación aditiva consigo mismo corresponde a $1+\gamma/2$, dado que $P(a \equiv b) = \gamma/2$. En consecuencia, la coancestría es la mitad de la relación aditiva, es decir, $1/2+\gamma/4$ y resulta de considerar los cuatro modos de muestrear *a* y *b* con reposición. Nótese que en todas las expresiones presentadas hasta aquí, si se toma $\gamma = 0$, producto de considerar que los individuos de la base no se encuentran relacionados entre sí, se obtiene para el individuo X una relación aditiva de 1.

A la hora de calcular las relaciones aditivas de generaciones recientes tomando en cuenta la falta de independencia entre los individuos base, Legarra et al., (2015) propusieron emplear las reglas del método tabular (Emik y Terril, 1949). En este escenario, sólo es necesario contar con el valor γ para incorporarlo en el cómputo de la matriz $\boldsymbol{A}$ de relaciones aditivas. El objetivo es generar una matriz $\boldsymbol{A}$ modificada producto de considerar a los individuos base relacionados entre sí, que denotaremos como $\boldsymbol{A}^{\gamma}$. Dicha matriz toma la siguiente forma:

$$\boldsymbol{A}^{\gamma} = \boldsymbol{A}\left(1-\frac{\gamma}{2}\right)+\boldsymbol{J}\gamma \qquad [2.4]$$

donde $\boldsymbol{A}$ es la matriz de relaciones aditivas clásica (asumiendo no relacionados a los individuos base) y $\boldsymbol{J}$ es una matriz de unos. Más adelantes, en la sección 2.2.1.3 se presenta un ejemplo de dicha matriz. Jacquard (1974, p. 169) propuso el mismo modo de cálculo, pero expresándolo en términos escalares y de coancestrías en lugar de relaciones aditivas del siguiente modo

$$\Phi_{\mathrm{T}} = \Phi_0 + \left(1-\Phi_0\right)\alpha \qquad [2.5]$$

En [2.5] el valor Φ_0 es el coeficiente de coancestría calculado empleando solo la información genealógica, Φ_{T} es el coeficiente de coancestría total considerando la información de genealogía conocida y las relaciones entre los individuos base de la población (α).

Por su parte, otros algoritmos desarrollados para el computo de la consanguinidad como ser el de Quaas (1976) y el de Meuwissen y Luo (1992) o el de Henderson (1976) para el cálculo de la matriz inversa y rala de relaciones aditivas, deben modificarse para considerar las relaciones entre individuos base (Legarra *et al.*, 2015). Dichas modificaciones son complejas y no es posible generalizarlos correctamente a varias poblaciones, como si lo es con los MF, tal como se describe más adelante.

2.2.1.2. Algoritmos para calcular las relaciones de parentesco al incorporar un metafundador

Un MF es un *pseudo*-individuo que puede ser considerado padre y madre de todos los animales de la población base (Legarra *et al.* 2015), tal como se representa en la Figura 2.3. De hecho, el MF (individuo 1) de la Figura 2.3 representa a la población ancestral de la Figura 2.1 y permite condensar la información de las relaciones entre los animales base.

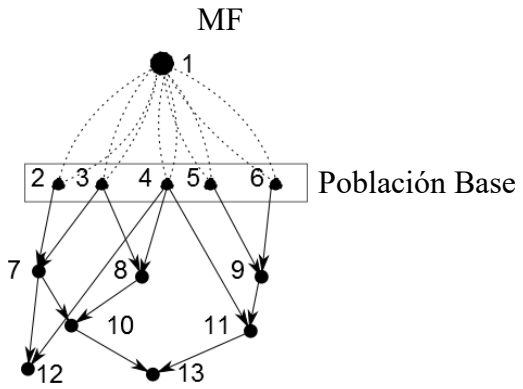


Figura 2.3. Esquema de una población en la que se incorpora un único MF (individuo 1). Las líneas punteadas representan relaciones imaginarias con dicho *pseudo*-individuo, las mismas permiten ilustrar la falta de independencia entre los individuos de la población base (2, 3, 4, 5 y 6). Tomada y modificada de Legarra *et al.* (2015).

En la Figura 2.3, el MF representa a un *pool* finito de gametas, del cual provienen los individuos que constituyen la población base (2 al 6). Si se toman dos gametas al azar con reposición de dicho *pool*, la relación entre dichas gametas es de $\gamma/2$ y, consecuentemente, el MF posee una relación consigo mismo de $A_{11} = \gamma$ y un "coeficiente de consanguinidad (*F*) individual" de $F_1 = A_{11} - 1 = \gamma - 1$, tomando en general valores negativos (Legarra *et al.* 2015). En este punto es necesario recordar que, si bien el MF se incorpora del mismo modo que cualquier otro individuo, el mismo es una figura ficticia que permite condensar las relaciones entre los animales base. De hecho, al incorporarlo, el elemento diagonal correspondiente toma valores inferiores a la unidad, situación que no se da para los individuos reales. En consecuencia, para el caso de los MF no aplica la interpretación tradicional del coeficiente *F*, que recordemos está dado en términos de probabilidad de IBD y por definición no puede tomar valores negativos. En el marco de los MF, hacemos referencia a dicho valor como "consanguinidad" pero es necesario volver a recalcar que su interpretación no es la tradicional. Ahora bien, en términos generales, la consanguinidad implica un desvío del equilibrio Hardy-Weinberg (Legarra *et al.* 2015). En consecuencia, al tomar en consideración el caso especial de los MF, el hecho de observar un *F* negativo se asocia a un exceso de heterocigosidad. En consecuencia, coeficientes negativos implican que, en la mayoría de los casos, las dos gametas son diferentes. Dicho de otro modo, que el *pool* del cual provienen es grande, lo que resulta lógico desde el punto de vista genético. Llevándolo al extremo tal como lo hacen Legarra *et al.* (2015), si se toma el caso de $\gamma = 0$ (y, en consecuencia, $F = -1$) , dicho valor sugiere que las dos gametas siempre son distintas (por descendencia) y no se encuentran relacionadas, es decir, que el pool del cual provienen es infinito, la heterocigosidad por descendencia es completa. En dicho caso, todos los individuos de la población base no están relacionados. Por el

contrario, si se toma $\gamma = 2$ (y, en consecuencia, $F = 1$) representa el caso opuesto. Es decir que las dos gametas tomadas al azar son siempre IBD, el *pool* del cual provienen sólo cuenta con una gameta, la homocigosidad es total y todos los individuos en la población base son idénticos y completamente consanguíneos. Nótese que ambos casos analizados son extremos e improbables en poblaciones reales pero resultan ilustrativos a la hora de interpretar los valores de *F* (o de γ, dada la relación que existe entre ambos) del MF y su implicancia en la población bajo estudio.

Para calcular las relaciones aditivas entre los individuos de una población cuando los animales de la base se encuentran relacionados entre sí, Legarra *et al.* (2015) propusieron emplear las reglas tradicionales del método tabular (Emik y Terril, 1949) tal como se describe a continuación. Estas sólo se ven parcialmente modificadas al incorporar un único MF. Resumiéndolas muy brevemente y sin entrar en detalles, en un primer paso se le asigna una relación aditiva de uno a todos los animales base consigo mismos. Nótese que esto se realiza bajo el supuesto clásico de que dichos individuos provienen de una población ancestral no relacionada. Luego para el resto de los individuos se aplican las siguientes fórmulas:

$$\begin{aligned} A_{ij} &= \tfrac{1}{2}\left(A_{dj} + A_{sj}\right) \\ A_{ii} &= 1 + \tfrac{1}{2}\left(A_{sd}\right) \end{aligned} \qquad [2.6]$$

donde *d* y *s* corresponden al padre y a la madre de *i*, respectivamente. Este último animal debe ser más joven que *j*. Ahora bien, para incluir un MF en el cómputo sólo es necesario un cambio sencillo. Simplemente basta con incluir una fila y columna para el MF y asignar al elemento A_{11} de la matriz el valor de γ (Legarra *et al.* 2015). De hecho, no es necesario modificar el resto de las reglas. Por ejemplo, si se considera la Figura 2.3, para el individuo 2

$$A_{22} = 1 + \tfrac{1}{2} A_{11} = 1 + \tfrac{\gamma}{2} \ ,$$

y para los individuos 1 y 2,

$$A_{12} = \tfrac{1}{2}\left(A_{11} + A_{11}\right) = \gamma \, .$$

Del mismo modo, para los individuos 2 y 3,

$$A_{23} = \tfrac{1}{2}\left(A_{12} + A_{12}\right) = \gamma \, .$$

De este modo puede calcularse fácilmente $\boldsymbol{A}^{\gamma}$ y en la siguiente sección se presenta un ejemplo pequeño.

2.2.1.3. Ejemplo de una matriz de relaciones con un metafundador

A continuación se ilustra el modo en que puede incorporarse un MF a cierta genealogía con el objeto de considerar las relaciones entre los individuos base. A tal fin, considérese la genealogía de la Figura 2.4 a continuación. La misma consiste de cuatro individuos reales (2, 3, 4 y 5) y un MF (1). Nótese que éste último se encuentra etiquetado de manera distinta (en letra negrita y cursiva) con el objetivo de recordar que corresponde a un *pseudo*-individuo.

Individuo	Padre	madre
1	0	0
2	***1***	***1***
3	***1***	***1***
4	2	3
5	2	4

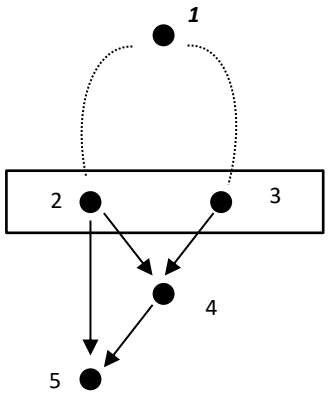


Figura 2.4. Ejemplo de una genealogía con cuatro individuos reales (2, 3, 4, y 5) **y un MF** (1)**.**

A los fines de ilustrar cómo incorporar un MF en el cálculo de $\boldsymbol{A}^{\gamma}$ se asume en este ejemplo $\gamma = A_{11} = 0,2$. Empleando el método tabular (Emik y Terril, 1949), $A_{22} = 1 + A_{11}/2 = 1,1$ y $A_{12} = 0,5(A_{1padre(2)} + A_{1madre(2)}) = 0,5(A_{11} + A_{11}) = 0,2$. Al continuar con las reglas de dicho método tabular, se obtiene la siguiente matriz $\boldsymbol{A}^{\gamma}$:

$$\boldsymbol{A}^{\gamma} = \begin{bmatrix} 0,200 & 0,200 & 0,200 & 0,200 & 0,200 \\ 0,200 & 1,100 & 0,200 & 0,650 & 0,875 \\ 0,200 & 0,200 & 1,100 & 0,650 & 0,425 \\ 0,200 & 0,650 & 0,650 & 1,100 & 0,875 \\ 0,200 & 0,875 & 0,425 & 0,875 & 1,325 \end{bmatrix}$$

La inversa ($\boldsymbol{A}^{\gamma-1}$) puede obtenerse invirtiendo γ y empleando las reglas de Henderson (1976):

$$\boldsymbol{A}^{\gamma-1} = \begin{bmatrix} 7,222 & -1,111 & -1,111 & 0,000 & 0,000 \\ -1,111 & 2,222 & 0,556 & -0,556 & -1,111 \\ -1,111 & 0,556 & 1,667 & -1,111 & 0,000 \\ 0,000 & -0,556 & -1,111 & 2,778 & -1,111 \\ 0,000 & -1,111 & 0,000 & -1,111 & 2,222 \end{bmatrix}$$

Ambas matrices pueden compararse con la $\boldsymbol{A}$ tradicional al fijar γ = 0. En este caso, el individuo 1 corresponde a un grupo de padres desconocidos y los elementos correspondientes de la matriz se fijaron a cero para ilustrar el efecto, tal como se presenta a continuación:

$$\boldsymbol{A} = \begin{bmatrix} 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 1,000 & 0,000 & 0,500 & 0,750 \\ 0,000 & 0,000 & 1,000 & 0,500 & 0,250 \\ 0,000 & 0,500 & 0,500 & 1,000 & 0,750 \\ 0,000 & 0,750 & 0,250 & 0,750 & 1,250 \end{bmatrix}$$

y la inversa de la matriz de relaciones incluyendo el grupo de padres desconocido (Quaas, 1988) es igual a

$$\boldsymbol{A}^{-1} = \begin{bmatrix} 2,000 & -1,000 & -1,000 & 0,000 & 0,000 \\ -1,000 & 2,000 & 0,500 & -0,500 & -1,000 \\ -1,000 & 0,500 & 1,500 & -1,000 & 0,000 \\ 0,000 & -0,500 & -1,000 & 2,500 & -1,000 \\ 0,000 & -1,000 & 0,000 & -1,000 & 2,000 \end{bmatrix}$$

2.2.2. Empleo de metafundadores en casos con múltiples poblaciones base

Hasta el momento se ha abordado el caso en que la población proviene de una única base. Ahora bien, se dan casos en los que las poblaciones actuales provienen de más de una ancestral y/o base tal como fue descripto, por ejemplo, en bovinos (Gibbs *et al.*, 2009). Para abordar este problema, Legarra *et al.* (2015) propusieron considerar dichos casos empleando más de un MF. A continuación se resume dicha propuesta y su implementación.

2.2.2.1. Algoritmos para calcular las relaciones entre y dentro de las poblaciones base con varios metafundadores

Legarra *et al.* (2015) propusieron un modo de abordar estas situaciones tomando como base el trabajo de VanRaden (1992) y su posterior implementación (VanRaden, 2011). La Figura 2.5 ilustra un caso posible de observar en poblaciones reales. Nótese que la población actual tiene dos bases (A y B) notablemente diferenciadas y ambas provienen, a su vez, de dos subpoblaciones ancestrales que se solapan parcialmente. La Figura 2.5 busca representar sólo uno de los casos posibles, tal vez de los más frecuentes, pero pueden encontrarse numerosas variantes, todas abordables empleando MF.

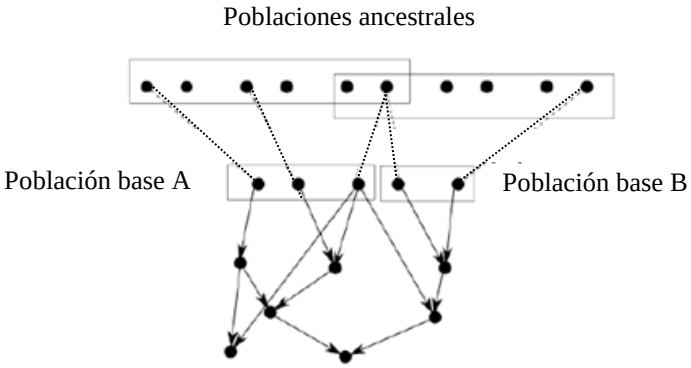


Figura 2.5. Esquema de un caso con dos poblaciones base provenientes de dos poblaciones ancestrales. Tomada y modificada de Legarra *et al.* (2015).

El modelo conceptual de los MF puede extenderse a varias poblaciones base con la posibilidad de que se solapen tal como se presenta en la Figura 2.5. En dicho caso, es necesario definir dos tipos de relaciones: entre las poblaciones *i* y *j* ($\gamma^{i,j}$) y dentro de poblaciones (γ^{i}, γ^{j}). Estas últimas se corresponden a la introducida previamente en la sección 2.2.1.1. En consecuencia, al momento de involucrar más de una población se obtiene una matriz como la siguiente:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{i} & \gamma^{i,j} & \cdots & \gamma^{i,n} \\ & \gamma^{j} & & \gamma^{j,n} \\ & & \ddots & \vdots \\ & \text{simétrica} & & \gamma^{n} \end{bmatrix}. \qquad [2.7]$$

La expresión [2.7] es general porque permite considerar *n* poblaciones. Nótese que la dimensión de $\boldsymbol{\Gamma}$ está dada por el número de poblaciones ancestrales a considerar (*n* x *n*). Ahora bien, es importante destacar la necesidad que la matriz $\boldsymbol{\Gamma}$ sea positiva definida porque, tal como se abordará más adelante, será necesario contar con su inversa (Legarra *et al.* 2015).

En relación con la interpretación de $\gamma^{i,j}$, dichos valores permiten cuantificar la relación entre ambas poblaciones base (*i* y *j*). Dicho de otro modo, cuando toman valores distintos de cero indican que ambas poblaciones base comparten ancestros comunes, tal como ocurre en la Figura 2.5. En la misma se solapan ambas poblaciones ancestrales (concretamente existen dos individuos que pertenecen a las dos poblaciones ancestrales que contribuyen a la formación de ambas poblaciones base).

Siguiendo la línea de razonamiento de la sección 2.2.1.1 y de Legarra *et al.* (2015), si se asume que la población *i* está compuesta por n_i gametas, la *j* por n_j y que ambas poblaciones comparten un total de n_{ij} gametas, los parámetros γ se definen del siguiente modo:

$$\gamma^{i} = 1/n_i; \qquad \gamma^{j} = 1/n_j; \qquad \gamma^{i,j} = n_{ij}/n_i\, n_j \qquad [2.8]$$

La interpretación γ^{i} y γ^{j} es análoga a la presentada en la sección 2.2.1.1. Por su parte, $\gamma^{i,j}$ involucra a la probabilidad que la gameta de *i* provenga de los ancestros de ambas poblaciones (n_{ij}/n_i); involucra además a la probabilidad que la gameta de *j* derive del mismo conjunto de individuos (n_{ij}/n_j) y, finalmente, a la probabilidad que ambas gametas sean iguales, dado que provienen del conjunto de individuos ancestrales de ambas poblaciones ($1/n_{ij}$). Es decir,

$$\gamma^{i,j} = P(\mathrm{g}_i \leftarrow \mathrm{g}_{ij})\, P(\mathrm{g}_j \leftarrow \mathrm{g}_{ij})\, P(\mathrm{g}_i = \mathrm{g}_j) = \frac{n_{ij}}{n_i}\frac{n_{ij}}{n_j}\frac{1}{n_{ij}} = \frac{n_{ij}}{n_i n_j} \qquad [2.9]$$

En el ejemplo de la Figura 2.5 tomado de Legarra *et al.* (2015), $n_{\mathrm{A}} = 6;\ n_{\mathrm{B}} = 4;\ n_{\mathrm{AB}} = 2$. Es decir que,

$$\gamma^{\mathrm{A}} = 1/n_{\mathrm{A}} = 1/6; \qquad \gamma^{\mathrm{B}} = 1/n_{\mathrm{B}} = 1/4; \qquad \gamma^{\mathrm{A,B}} = 2/24 = 1/12$$

En consecuencia, la matriz $\boldsymbol{\Gamma}$ será de orden 2×2 y tendrá la siguiente forma:

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1/6 & 1/12 \\ 1/12 & 1/4 \end{bmatrix}$$

Legarra *et al.* (2015) propusieron considerar a cada una de las poblaciones ancestrales como un MF. Es decir que los elementos de $\boldsymbol{\Gamma}$ corresponden a las relaciones entre y dentro de MF, valores que pueden obtenerse a partir del conocimiento de la historia de las poblaciones involucradas (por ejemplo, si divergieron hace ciertas generaciones), o pueden inferirse empleando información genómica, tal como se describirá más adelante. Consecuentemente, aplicando la noción de los MF en el caso representado en la Figura 2.5, sería necesario considerar dos *pseudo*-individuos. Puede esquematizarse el mismo caso de la Figura 2.5 en términos de MF obteniendo la Figura 2.6 que se presenta a continuación.

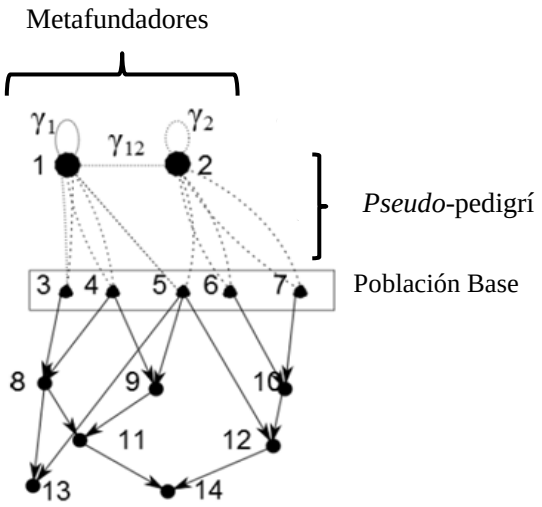


Figura 2.6. Esquema del caso con dos poblaciones base provenientes de dos ancestrales en el que se incorporan MF (individuos 1 y 2). Las líneas punteadas representan el *pseudo*-pedigrí (relaciones con los *pseudo*-individuos, entre ellos y consigo mismos). Tomada y modificada de Legarra *et al.* (2015).

De modo análogo al presentado para el caso de un MF, incluir varios MF da lugar a una matriz de relaciones $\boldsymbol{A}^{\boldsymbol{\Gamma}}$. Para calcularla Legarra *et al.* (2015) propusieron emplear los algoritmos mencionados anteriormente para el caso con un único MF (sección 2.2.1.2) y extenderlos para incluir múltiples MF. Concretamente, para emplear las reglas del método tabular (Emik y Terril, 1949) el primer paso es generar la matriz $\boldsymbol{\Gamma}$ con las relaciones entre los MF y, luego, aplicar las fórmulas propuestas originalmente. Ahora bien, para calcular la inversa de dicha matriz $\left(\boldsymbol{A}^{\boldsymbol{\Gamma}-1}\right)$ de modo directo es necesario, primero, invertir $\boldsymbol{\Gamma}$ (de aquí la importancia que sea positiva definida). Esto permite crear, en una primera instancia, una pequeña submatriz de $\boldsymbol{A}^{\boldsymbol{\Gamma}-1}$ y luego emplear las reglas de Henderson (1976) con los

elementos $\boldsymbol{D}_{ii}$ de todos los individuos modificados de acuerdo con la relación consigo mismo de los MF (γ^{i}), tal como lo describen con mayor grado de detalle Legarra *et al.* (2015).

2.2.2.2. Ejemplo

Presentaremos a continuación un ejemplo que ilustra el empleo de dos MF. El mismo fue adaptado de Legarra *et al.* (2015) y descripto en la Figura 2.6. Si se toman los 12 individuos reales (aquellos indicados con los números 3 al 14) y se calculan las relaciones aditivas entre ellos previo a la incorporación de los MF, se obtiene la matriz $\boldsymbol{A}$ tradicional. Esta conlleva el supuesto que los animales base (individuos 3 al 7) no se encuentran emparentados. Se indican en sombreado ciertas relaciones con el objeto de prestar especial atención a cómo varían a lo largo de los escenarios propuestos. En este primer caso la matriz $\boldsymbol{A}$ es la siguiente:

```
      [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10][,11][,12][,13] [,14]
[3,]  1,00 0,00 0,00 0,00 0,00 0,50 0,00 0,00 0,25 0,00  0,25  0,12
[4,]  0,00 1,00 0,00 0,00 0,00 0,50 0,50 0,00 0,50 0,00  0,25  0,25
[5,]  0,00 0,00 1,00 0,00 0,00 0,00 0,50 0,00 0,25 0,50  0,50  0,38
[6,]  0,00 0,00 0,00 1,00 0,00 0,00 0,00 0,50 0,00 0,25  0,00  0,12
[7,]  0,00 0,00 0,00 0,00 1,00 0,00 0,00 0,50 0,00 0,25  0,00  0,12
[8,]  0,50 0,50 0,00 0,00 0,00 1,00 0,25 0,00 0,62 0,00  0,50  0,31
[9,]  0,00 0,50 0,50 0,00 0,00 0,25 1,00 0,00 0,62 0,25  0,38  0,44
[10,] 0,00 0,00 0,00 0,50 0,50 0,00 0,00 1,00 0,00 0,50  0,00  0,25
[11,] 0,25 0,50 0,25 0,00 0,00 0,62 0,62 0,00 1,12 0,12  0,44  0,62
[12,] 0,00 0,00 0,50 0,25 0,25 0,00 0,25 0,50 0,12 1,00  0,25  0,56
[13,] 0,25 0,25 0,50 0,00 0,00 0,50 0,38 0,00 0,44 0,25  1,00  0,34
[14,] 0,12 0,25 0,38 0,12 0,12 0,31 0,44 0,25 0,62 0,56  0,34  1,06
```

Ahora bien, yendo un paso más allá, es posible recalcularla considerando que los animales base provienen de poblaciones ancestrales de tamaño pequeño y, en consecuencia, se encuentran relacionados entre sí. A tal fin es necesario incorporar los MF como se discutió en la sección precedente. Dado que en este caso existen dos poblaciones, será necesario incorporar dos MF (*pseudo*-individuos 1 y 2). A continuación se exploran dos escenarios: el primero considera dos poblaciones base no relacionadas entré sí, no así los individuos de cada una; el segundo incorpora una relación distinta de cero entre ambas poblaciones base.

Considérese, en un primer escenario, que $\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{A} & \gamma^{A,B} \\ \gamma^{B,A} & \gamma^{B} \end{bmatrix} = \begin{bmatrix} 0,1 & 0 \\ 0 & 0,2 \end{bmatrix}$, es decir que ambas poblaciones no se encuentran relacionadas ancestralmente, pero existen relaciones entre los individuos fundadores de cada población. En este caso, la matriz $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ es la siguiente:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
[1,]	0,10	0,00	0,10	0,10	0,05	0,00	0,00	0,10	0,08	0,00	0,09	0,02	0,08	0,06
[2,]	0,00	0,20	0,00	0,00	0,10	0,20	0,20	0,00	0,05	0,20	0,02	0,15	0,05	0,09
[3,]	0,10	0,00	1,05	0,10	0,05	0,00	0,00	0,58	0,08	0,00	0,33	0,02	0,31	0,18
[4,]	0,10	0,00	0,10	1,05	0,05	0,00	0,00	0,58	0,55	0,00	0,56	0,02	0,31	0,29
[5,]	0,05	0,10	0,05	0,05	1,00	0,10	0,10	0,05	0,52	0,10	0,29	0,55	0,52	0,42
[6,]	0,00	0,20	0,00	0,00	0,10	1,10	0,20	0,00	0,05	0,65	0,02	0,38	0,05	0,20
[7,]	0,00	0,20	0,00	0,00	0,10	0,20	1,10	0,00	0,05	0,65	0,02	0,38	0,05	0,20
[8,]	0,10	0,00	0,58	0,58	0,05	0,00	0,00	1,05	0,31	0,00	0,68	0,02	0,55	0,35
[9,]	0,08	0,05	0,08	0,55	0,52	0,05	0,05	0,31	1,02	0,05	0,67	0,29	0,42	0,48
[10,]	0,00	0,20	0,00	0,00	0,10	0,65	0,65	0,00	0,05	1,10	0,02	0,60	0,05	0,31
[11,]	0,09	0,02	0,33	0,56	0,29	0,02	0,02	0,68	0,67	0,02	1,16	0,16	0,48	0,66
[12,]	0,02	0,15	0,02	0,02	0,55	0,38	0,38	0,02	0,29	0,60	0,16	1,05	0,29	0,60
[13,]	0,08	0,05	0,31	0,31	0,52	0,05	0,05	0,55	0,42	0,05	0,48	0,29	1,02	0,39
[14,]	0,06	0,09	0,18	0,29	0,42	0,20	0,20	0,35	0,48	0,31	0,66	0,60	0,39	1,08

Nótese que todas las relaciones dentro de cada raza (A: individuos 8, 9, 11, 13 y B: individuos 10 y 12) aumentaron. Esto se debe a que ahora los animales de la base se encuentran emparentados. Por su parte, la relación entre los individuos 8 y 10 (cada uno de diferente población) se mantuvo en cero debido a que aquí no se contempló la relación entre individuos de ambas poblaciones base.

En un segundo escenario, se considera además que ambas poblaciones base se encuentren relacionadas entre sí. Como ejemplo, considérese a $\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{A} & \gamma^{A,B} \\ \gamma^{B,A} & \gamma^{B} \end{bmatrix} = \begin{bmatrix} 0,10 & 0,05 \\ 0,05 & 0,20 \end{bmatrix}$. En este caso la matriz $\boldsymbol{A}^{\boldsymbol{\Gamma}}$ es igual a:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
[1,]	0,10	0,05	0,10	0,10	0,08	0,05	0,05	0,10	0,09	0,05	0,09	0,06	0,09	0,08
[2,]	0,05	0,20	0,05	0,05	0,12	0,20	0,20	0,05	0,09	0,20	0,07	0,16	0,09	0,12
[3,]	0,10	0,05	1,05	0,10	0,08	0,05	0,05	0,58	0,09	0,05	0,33	0,06	0,33	0,20
[4,]	0,10	0,05	0,10	1,05	0,08	0,05	0,05	0,58	0,56	0,05	0,57	0,06	0,33	0,32
[5,]	0,08	0,12	0,08	0,08	1,02	0,12	0,12	0,08	0,55	0,12	0,31	0,57	0,55	0,44
[6,]	0,05	0,20	0,05	0,05	0,12	1,10	0,20	0,05	0,09	0,65	0,07	0,39	0,09	0,23
[7,]	0,05	0,20	0,05	0,05	0,12	0,20	1,10	0,05	0,09	0,65	0,07	0,39	0,09	0,23
[8,]	0,10	0,05	0,58	0,58	0,08	0,05	0,05	1,05	0,33	0,05	0,69	0,06	0,56	0,38
[9,]	0,09	0,09	0,09	0,56	0,55	0,09	0,09	0,33	1,04	0,09	0,68	0,32	0,44	0,50
[10,]	0,05	0,20	0,05	0,05	0,12	0,65	0,65	0,05	0,09	1,10	0,07	0,61	0,09	0,34
[11,]	0,09	0,07	0,33	0,57	0,31	0,07	0,07	0,69	0,68	0,07	1,16	0,19	0,50	0,68
[12,]	0,06	0,16	0,06	0,06	0,57	0,39	0,39	0,06	0,32	0,61	0,19	1,06	0,32	0,63
[13,]	0,09	0,09	0,33	0,33	0,55	0,09	0,09	0,56	0,44	0,09	0,50	0,32	1,04	0,41
[14,]	0,08	0,12	0,20	0,32	0,44	0,23	0,23	0,38	0,50	0,34	0,68	0,63	0,41	1,10

Nótese que la relación entre los animales 8 y 10 deja de ser cero, lo que impacta directamente en el coeficiente de consanguinidad del animal 14, el cual aumenta.

Puede plantearse también una variante al escenario anterior en el que se incrementa el valor de γ^{B} a 0,30 para ilustrar el impacto que tiene el valor de γ^{B} sobre los elementos de la matriz $\boldsymbol{A}^{\boldsymbol{\Gamma}}$. Esta ahora es igual a

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{A} & \gamma^{A,B} \\ \gamma^{B,A} & \gamma^{B} \end{bmatrix} = \begin{bmatrix} 0,10 & 0,05 \\ 0,05 & 0,30 \end{bmatrix}$$

La matriz A^{Γ} resultante es

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
[1,]	0,10	0,05	0,10	0,10	0,08	0,05	0,05	0,10	0,09	0,05	0,09	0,06	0,09	0,08
[2,]	0,05	0,30	0,05	0,05	0,18	0,30	0,30	0,05	0,11	0,30	0,08	0,24	0,11	0,16
[3,]	0,10	0,05	1,05	0,10	0,08	0,05	0,05	0,58	0,09	0,05	0,33	0,06	0,33	0,20
[4,]	0,10	0,05	0,10	1,05	0,08	0,05	0,05	0,58	0,56	0,05	0,57	0,06	0,33	0,32
[5,]	0,08	0,18	0,08	0,08	1,02	0,18	0,18	0,08	0,55	0,18	0,31	0,60	0,55	0,46
[6,]	0,05	0,30	0,05	0,05	0,18	1,15	0,30	0,05	0,11	0,72	0,08	0,45	0,11	0,27
[7,]	0,05	0,30	0,05	0,05	0,18	0,30	1,15	0,05	0,11	0,72	0,08	0,45	0,11	0,27
[8,]	0,10	0,05	0,58	0,58	0,08	0,05	0,05	1,05	0,33	0,05	0,69	0,06	0,56	0,38
[9,]	0,09	0,11	0,09	0,56	0,55	0,11	0,11	0,33	1,04	0,11	0,68	0,33	0,44	0,51
[10,]	0,05	0,30	0,05	0,05	0,18	0,72	0,72	0,05	0,11	1,15	0,08	0,66	0,11	0,37
[11,]	0,09	0,08	0,33	0,57	0,31	0,08	0,08	0,69	0,68	0,08	1,16	0,20	0,50	0,68
[12,]	0,06	0,24	0,06	0,06	0,60	0,45	0,45	0,06	0,33	0,66	0,20	1,09	0,33	0,64
[13,]	0,09	0,11	0,33	0,33	0,55	0,11	0,11	0,56	0,44	0,11	0,50	0,33	1,04	0,42
[14,]	0,08	0,16	0,20	0,32	0,46	0,27	0,27	0,38	0,51	0,37	0,68	0,64	0,42	1,10

El análisis de los elementos de esta última matriz muestra que sólo aumentaron los elementos correspondientes a los individuos de la población B, aquellos que involucran a los animales 10 y 12.

Finalmente, y a modo de verificación, se considera la situación en que los MF no se encuentran relacionados entre sí ni consigo mismos. Dicho de otro modo, este caso responde al supuesto tradicional de que los individuos base no están emparentados. En este caso

$$\Gamma = \begin{bmatrix} \gamma^{A} & \gamma^{A,B} \\ \gamma^{B,A} & \gamma^{B} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Se obtiene la siguiente matriz:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
[1,]	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
[2,]	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
[3,]	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,50	0,00	0,00	0,25	0,00	0,25	0,12
[4,]	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,50	0,50	0,00	0,50	0,00	0,25	0,25
[5,]	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,50	0,00	0,25	0,50	0,50	0,38
[6,]	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,50	0,00	0,25	0,00	0,12
[7,]	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,50	0,00	0,25	0,00	0,12
[8,]	0,00	0,00	0,50	0,50	0,00	0,00	0,00	1,00	0,25	0,00	0,62	0,00	0,50	0,31
[9,]	0,00	0,00	0,00	0,50	0,50	0,00	0,00	0,25	1,00	0,00	0,62	0,25	0,38	0,44
[10,]	0,00	0,00	0,00	0,00	0,00	0,50	0,50	0,00	0,00	1,00	0,00	0,50	0,00	0,25
[11,]	0,00	0,00	0,25	0,50	0,25	0,00	0,00	0,62	0,62	0,00	1,12	0,12	0,44	0,62
[12,]	0,00	0,00	0,00	0,00	0,50	0,25	0,25	0,00	0,25	0,50	0,12	1,00	0,25	0,56
[13,]	0,00	0,00	0,25	0,25	0,50	0,00	0,00	0,50	0,38	0,00	0,44	0,25	1,00	0,34
[14,]	0,00	0,00	0,12	0,25	0,38	0,12	0,12	0,31	0,44	0,25	0,62	0,56	0,34	1,06

Como era de esperar, se obtiene la matriz A tradicional para los individuos reales (3 al 14) e idéntica a la presentada al inicio de la sección.

2.2.3. Estimación de las relaciones ancestrales de los metafundadores empleando información genómica por máxima verosimilitud (Christensen, 2012) y por el método de momentos (Legarra *et al.*, 2015)

En esta sección se presentan los métodos disponibles hasta el momento para estimar los parámetros γ. El objetivo de esta sección es describir los algoritmos disponibles, sus características, ventajas y desventajas con el objeto de proponer métodos alternativos más sencillos y evaluar su desempeño en el capítulo 3 de esta tesis.

Hasta el momento se han presentado los fundamentos teóricos con ejemplos simples que asumían conocido el pedigrí y el *pseudo*-pedigrí entre los MF. Ahora bien, este último es desconocido en escenarios reales y, en consecuencia, las relaciones entre y dentro de los fundadores no pueden inferirse a partir de la información genealógica. Para resolver este inconveniente, se propusieron metodología para estimarlas empleando la información de marcadores moleculares refiriéndolas a una base genética definida de acuerdo con las relaciones genómicas (Christensen 2012 y Legarra et *al.*, 2015). A tal fin, se propusieron dos modos de estimar los parámetros γ: i) método de máxima verosimilitud (Christensen, 2012) y ii) método de momentos (MM - Legarra *et al.*, 2015). Ambas metodologías se describen a continuación.

Al momento de introducir el parámetro γ, Christensen (2012) propuso también un modo de estimarlo empleando información genómica: el método de máxima verosimilitud. Dado que los genotipos de los marcadores poseen herencia mendeliana, la covarianza entre los genotipos de dos individuos es igual a su relación genómica. El método propuesto por Christensen (2012) para estimar γ en una única población se fundamenta en dicho principio. Tal como lo presentó Legarra *et al.* (2015), en un primer paso integró la verosimilitud sobre las frecuencias alélicas, lo que resultó en el empleo de frecuencias alélicas de 0,5 como referencia (los genotipos de la matriz $\boldsymbol{Z}$ codificados como {−1, 0, 1}). Al asumir normalidad multivariada para los genotipos de los marcadores (matriz $\boldsymbol{Z}$), la verosimilitud condicional a γ y *s* puede expresarse del siguiente modo

$$\log \mathrm{p}\left(\boldsymbol{Z}\middle|\gamma, s\right) = \mathrm{const} - \frac{p\,n}{2}\log\left(s\right) - \frac{p}{2}\log\left|\boldsymbol{A}_{22}^{\gamma}\right| - \frac{p}{2s}\mathrm{tr}\left(\boldsymbol{A}_{22}^{\gamma-1}\boldsymbol{Z}\boldsymbol{Z}'\right) \qquad [2.10]$$

donde *n* es el número de individuos genotipados y $\boldsymbol{A}_{22}^{\gamma}$ es la submatriz de $\boldsymbol{A}^{\gamma}$ correspondiente a los individuos genotipados. Por su parte, el parámetro *s* corresponde a una medida de heterocigosidad en la población genotipada. La extensión de esta verosimilitud a múltiples poblaciones con diferentes parámetros γ en $\boldsymbol{\Gamma}$ es directa y se presenta con mayor grado de detalle en Legarra et al. (2015). El procedimiento puede completarse agregando una distribución *a priori* para γ o $\boldsymbol{\Gamma}$ y empleando un estimador bayesiano en vez de máxima verosimilitud. De todos modos, en ningún caso es posible

factorizar hacia afuera a γ o $\boldsymbol{\Gamma}$. La verosimilitud [2.10] no tiene una forma explícita que permita la maximización de manera analítica, debiendo optimizarse mediante un proceso de búsqueda numérico-estocástico como por ejemplo Monte Carlo (Legarra *et al.* 2015).

Por otro lado, Legarra *et al.* (2015) propuso el método de momentos (MM) como una alternativa más sencilla a la presentada por Christensen (2012). MM combina estadísticos resumen de las matrices $\boldsymbol{A}_{22}^{\Gamma}$ y $\boldsymbol{G}$ (VanRaden *et al.*, 2008; Vitezica *et al.*, 2011; Christensen *et al.*, 2012). Combinar estos estadísticos permite forzar la equivalencia entre los cambios esperados de la media y la varianza bajo deriva genética (Vitezica *et al.*, 2011; Christensen *et al.*, 2012) para la población descripta, ya sea por la genealogía o por las matrices de relaciones genómicas (Legarra *et al.* 2015). Con tal fin Legarra *et al.* (2015) consideraron tres situaciones posibles: i) una única población; ii) múltiples poblaciones puras; iii) poblaciones con animales puros y cruzas.

Para el caso i) en el que se cuenta con una única población, es necesario estimar dos parámetros: γ y *s*. Legarra *et al.* (2015) establecieron un sistema de ecuaciones de orden 2 x 2 y obtuvieron las siguientes expresiones como soluciones:

$$\gamma = \frac{\overline{\boldsymbol{ZZ'}}/s - \overline{\boldsymbol{A}_{22}}}{1 - \overline{\boldsymbol{A}_{22}}/2} \qquad [2.11]$$

$$s = \frac{\overline{\text{diag}}(\boldsymbol{ZZ'})(1 - \overline{\boldsymbol{A}_{22}}/2) - \overline{\boldsymbol{ZZ'}}(1 - \overline{\text{diag}}(\boldsymbol{A}_{22})/2)}{\overline{\text{diag}}(\boldsymbol{A}_{22}) - \overline{\boldsymbol{A}_{22}}} \qquad [2.12]$$

Nótese entonces que es posible emplear [2.11] y [2.12] para estimar ambos parámetros de interés. La ventaja de este método es que sólo requiere valores tomados a partir de las matrices $\boldsymbol{A}_{22}$ y $\boldsymbol{G}$, que pueden obtenerse fácilmente, incluso sin necesidad de construirlas.

Ahora bien, en el caso de contar con más de una población de razas puras, es posible extender la metodología previamente presentada de modo sencillo tal como lo presentan Legarra *et al.* (2015). Por ejemplo, y por simplicidad, en el caso de contar con dos poblaciones serán cuatro los parámetros a estimar: $\gamma_A, \gamma_B, \gamma_{AB}$ y s. De hecho, la manera de estimar las relaciones entre razas es similar a la propuesta por VanRaden *et al.* (2011), pero la diferencia con esta radica en el modo de escalar las matrices genómicas, que en el caso de Legarra *et al.* (2015) emplea *s*. Es posible utilizar este método con más de dos poblaciones puras y, tal como sucede con el caso de una única población, sólo es necesario contar con estadísticos de las matrices de relaciones. Para el caso de poblaciones con animales puros y cruzas, Legarra *et al.* (2015) propusieron un método similar al presentado por Harris y Johnson (2010).

2.2.4. Combinación de relaciones genealógicas con metafundadores y relaciones genómicas

El método ssGBLUP para evaluaciones genómicas (Aguilar *et al.*, 2010; Christensen y Lund 2010; Legarra *et al.*, 2014) permite combinar la información genómica con la genealógica. A tal fin se emplea la matriz $\boldsymbol{H}$ (Legarra *et al.*, 2009; Christensen y Lund 2010) cuya estructura se presentó en la sección 2.1 de este capítulo (expresión [2.1]). El desarrollo algebraico de la matriz $\boldsymbol{H}$ asume que las frecuencias alélicas de la población base son conocidas o, lo que es equivalente, que la media y la varianza de la población no cambia a lo largo del tiempo (Legarra *et al.* 2015). Este supuesto no se cumple en poblaciones pequeñas, pedigríes profundos o cuando hay selección, situaciones que suelen darse en la gran mayoría de los casos en poblaciones reales. Con el objetivo de abordar esta problemática Vitezica *et al.* (2011), Christensen *et al.* (2012), entre otros, propusieron ajustes que permiten modificar las relaciones genómicas de modo tal que la base genética considerada sea la misma que aquella tomada para las relaciones de pedigrí. De todos modos, estos ajustes no permiten considerar la estructura del pedigrí y su generalización a los cruzamientos entre líneas o razas no se encuentra completamente desarrollada ni comprendida en su totalidad (Legarra *et al.*, 2015).

Christensen (2012) propuso un ajuste alternativo: compatibilizar las relaciones de pedigrí con las genómicas en lugar de lo opuesto como era lo usual hasta el momento. El autor mostró que al marginalizar las frecuencias alélicas de la verosimilitud conjunta, se observa una población base relacionada y sugirió estimar γ y s usando máxima verosimilitud (Legarra *et al.*, 2015). Tal como se mencionó anteriormente, el desarrollo teórico presentado por Legarra *et al.* (2015) en relación a los MF se fundamenta en las bases propuestas por Christensen (2012). El uso de MF con relaciones $\boldsymbol{\Gamma}$ permite compatibilizar las relaciones calculadas condicionales al pedigrí, las genómicas y la consanguinidad (Powell *et al.*, 2010; Vitezica *et al.*, 2011). La homocigosidad puede considerarse como una desviación del equilibrio Hardy-Weinberg (Wright, 1922). Estas desviaciones no son fácilmente medibles debido a que dependen de las frecuencias alélicas consideradas que varían con el tiempo. Trabajar sobre el supuesto de fundadores no relacionados implica asumir que todos los alelos fundadores son diferentes entre sí. Esto no se sostiene al momento de contrastarlo con lo que ocurre en la realidad al disponer de información de marcadores moleculares. En cambio, al asumir frecuencias alélicas de 0,5 para la construcción de $\boldsymbol{G}$, la referencia es constante y no existen ambigüedades. Las genealogías extendidas considerando a los MF permiten conciliar automáticamente las relaciones calculadas por pedigrí y por información molecular empleando estimaciones de $\boldsymbol{\Gamma}$ y s basadas en los marcadores. En particular, la inversa de la matriz $\boldsymbol{H}$, al considerar MF es

$$\boldsymbol{H}^{\boldsymbol{\Gamma}-1} = \boldsymbol{A}^{\boldsymbol{\Gamma}-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-1} \end{bmatrix} \qquad [2.13]$$

Dicha matriz puede introducirse en las ecuaciones de modelo mixto (MME) del ssGBLUP (Legarra *et al.* 2015).

Por otro lado, Legarra *et al.* (2015) mostraron la relación teórica que existe entre los MF con los UPG o grupos genéticos (Thompson 1979; Quaas 1988). Estos últimos permiten estimar los valores esperados de distintas bases genéticas dentro de una misma población, lo cual es necesario cuando el apareamiento no es aleatorio, por ejemplo al incorporar animales importados, o cuando el pedigrí es incompleto de manera no aleatoria por causa de la selección (y no están reportados los individuos no selectos). Tal como señaló Kennedy (1991) la formulación tradicional de los grupos genéticos no toma en cuenta la consanguinidad o la deriva. La propuesta de los MF realizada por Legarra *et al.* (2015) es una generalización de los grupos genéticos que permite tomar en consideración la consanguinidad, la deriva y las relaciones entre los distintos grupos. Esta generalización permite superar los problemas mencionados por Misztal *et al.* (2013), quienes notaron que la incorporación de los UPG a los métodos *single-step* requería aproximaciones en el armado de la matriz $\boldsymbol{H}$.

2.2.5. Relación entre metafundadores y las frecuencias alélicas de la población base

En esta sección se aborda primero el caso con una única población y luego se procede a describir el escenario con más de una.

Sea $\boldsymbol{M}$ una matriz de genotipos codificada como {0, 1, 2} de orden *n* (número de animales genotipados) x *m* (número de marcadores) y la matriz de relaciones genómicas

$$\boldsymbol{G} = \left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' \Big/ s \qquad [2.14]$$

Donde $\boldsymbol{J}$ (*n* x *m*) es una matriz cuyos elementos toman el valor uno y los alelos de referencia se definen al azar con el objetivo de que la frecuencia alélica esperada sea de 0,5 (Christensen, 2012). Dicho de otro modo, en este caso $\boldsymbol{Z} = \left(\boldsymbol{M} - \boldsymbol{J}\right)$ y sus elementos toman valores {−1, 0 ,1} para cada genotipo. El orden de la matriz $\boldsymbol{G}$ en [2.14] es *n* x *n*. Para el caso de una única población, γ corresponde a la relación entre los fundadores de pedigrí, tal como se describió en las secciones precedentes. Por su parte, el parámetro *s*, también definido anteriormente, corresponde a una medida de la heterocigosidad de los marcadores en la población. La relación ancestral γ explica relaciones genómicas en [2.14] que no son capturadas por la genealogía (por ejemplo por tratarse de individuos que en el pedigrí no se encuentran relacionados dado que la información del mismo no es completa).

A continuación se muestra la relación, por un lado, entre γ y la varianza de las frecuencias alélicas de la población base y, por otro, entre *s* y el número de marcadores empleados en el análisis.

2.2.5.1. Derivación analítica de los parámetros γ y s (caso con una única población)

Esta sección y las dos siguientes constituyen las contribuciones originales del capítulo. En la presente sección se demuestra que los parámetros de relaciones ancestrales (γ) son proporcionales a las covarianzas de las frecuencias alélicas base de la población, tal como ocurre con el índice de fijación F_{ST} de Wright (1943), que será tratado con mayor grado de detalle en la siguiente sección. Establecer esta relación entre el parámetro **γ** y las varianzas de las frecuencias alélicas en la población base, constituye la base fundamental para estimar el parámetro empleando la varianza estimada de las frecuencias alélicas base, mediante información genómica. Este último aspecto será abordado en detalle en el próximo capítulo, pero cabe destacar su importancia dado su impacto potencial en la implementación sencilla de los MF en evaluaciones genómicas en distintas especies domésticas. Además, se demuestra la relación entre el parámetro *s* y el número de marcadores, lo que permite obtener una estimación de modo muy sencillo y sin costo computacional alguno.

Para una población particular, la estructura de (co)varianza es función de dos parámetros η_1 *y* η_2 del siguiente modo*:*

$$\gamma = \frac{4\eta_1}{2\eta_1 + \eta_2} \quad \text{y} \quad s = n\left(2\eta_1 + \eta_2\right) \qquad [2.15]$$

Ambas dependen de las frecuencias alélicas tal como se presenta en el apéndice de Christensen (2012) y *n* corresponde al número de marcadores. Si p_j corresponde a las frecuencias alélicas de los *j* = 1,…,*n* loci, estos parámetros son independientes de *j* y son iguales a:

$$\eta_1 = \mathrm{Var}\left(p_j\right) \quad \text{y} \quad \eta_2 = \mathrm{E}\left(2\, p_j\, q_j\right), \qquad [2.16]$$

siendo $q = 1 - p$.

Ahora bien, sabiendo que

$$\mathrm{E}\left(p\, q\right) = \mathrm{E}\left(p\left(1-p\right)\right) = \mathrm{E}\left(p\right) - \mathrm{E}\left(p^2\right) \qquad [2.17]$$

Y dado que

$$\mathrm{Var}(p) = \mathrm{E}(p^2) - \mathrm{E}(p)^2, \qquad [2.18]$$

se obtiene

$$\mathrm{E}(p^2) = \mathrm{Var}(p) + \mathrm{E}(p)^2 . \qquad [2.19]$$

Tenemos también $\mathrm{E}(q) = 1 - \mathrm{E}(p)$. Al sustituir $\mathrm{E}(p)^2$ de la expresión [2.19] en [2.17] se obtiene lo siguiente:

$$\begin{aligned} \mathrm{E}(pq) &= \mathrm{E}(p) - \mathrm{Var}(p) - \mathrm{E}(p)^2 \\ &= \mathrm{E}(p)(1 - \mathrm{E}(p)) - \mathrm{Var}(p) \\ &= \mathrm{E}(p)\mathrm{E}(q) - \mathrm{Var}(p) \end{aligned} \qquad [2.20]$$

Si los marcadores son bialélicos y se los toma aleatoriamente como referencia, entonces

$$\mathrm{E}(p) = \mathrm{E}(q) = 0,5 . \qquad [2.21]$$

En consecuencia, de la expresión [2.20] se obtiene

$$\mathrm{E}(pq) = 0,25 - \mathrm{Var}(p) . \qquad [2.22]$$

De aquí se obtiene

$$2\eta_1 + \eta_2 = 2\mathrm{Var}(p_j) + 0,5 - 2\mathrm{Var}(p_j) = 0,5 \qquad [2.23]$$

Y, en consecuencia,

$$s = n(2\eta_1 + \eta_2) = 0,5n = \frac{n}{2} , \qquad [2.24]$$

que, dicho en otras palabras, *s* corresponde a la mitad del número de marcadores empleados en el análisis. Además

$$\gamma = \frac{4\eta_1}{2\eta_1 + \eta_2} = \frac{4\eta_1}{0,5} = 8\mathrm{Var}(p_j) = 8\sigma_p^2 \qquad [2.25]$$

De modo tal que para una única población, γ es ocho veces la varianza de las frecuencias alélicas en la población base (descripta por Cockerham, 1969). Es importante destacar que $\mathrm{Var}(p_j) = \sigma_p^2$, con el objetivo de mostrar que σ_p^2 (y también γ) es un parámetro, la varianza de las frecuencias alélicas de los marcadores (Toro *et al.*, 2011; Christensen, 2012; Wright, 1931; Crow y Kimura, 1970). En el caso de *s*, puede considerárselo como equivalente a la heterocigosidad cuando todos los marcadores poseen una frecuencia alélica de 0,5, es decir el máximo valor posible de heterocigosidad. El resultado presentado en la

expresión [2.25] será de gran importancia de aquí en adelante, ya que sobre esta base se propondrán métodos sencillos para estimar γ en el próximo capítulo.

Ahora bien, dada la relación establecida entre *s* y el número de marcadores en [2.24], la matriz de relaciones genómicas de [2.14] puede expresarse como

$$\boldsymbol{G} = 2\left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' \Big/ n \qquad [2.26]$$

Nótese que la matriz en [2.26] es similar a aquella de relaciones IBS con la siguiente forma:

$$\boldsymbol{G}_{\text{IBS}} = \left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' \Big/ n + \boldsymbol{11}' \qquad [2.27]$$

De modo que $\boldsymbol{G}_{\text{IBS}} = \frac{1}{2}\boldsymbol{G} + \boldsymbol{11}'$, tal como se presenta en el apéndice.

2.2.5.2. Relaciones entre metafundadores para casos con más de una población

De modo análogo, la relación entre dos MF b y b' está dada por $\gamma_{b,b'} = 8\text{Cov}(p_{b,j}, p_{b',j}) = 8\sigma_{p_b,p_{b'}}$, es decir la covarianza entre loci entre las frecuencias alélicas de dos poblaciones b y b'. Nótese entonces que la relación (en este caso entre dos poblaciones) está dada por la covarianza entre el contenido génico en un locus. Christensen *et al.* (2015) muestran dicha relación de modo implícito. Por su parte, Cockerham (1969) y Robertson (1975) interpretaron a $4\sigma_{p_b,p_{b'}}$ como la coancestría entre dos poblaciones y Fariello *et al.* (2013) empleó $\sigma_{p_b,p_{b'}}$ para describir la divergencia de las poblaciones. De hecho, se han propuesto diversas medidas de la distancia genética entre poblaciones y la gran mayoría involucran un término que se relaciona, directa o indirectamente con $\sigma_{p_b,p_{b'}}$.

2.2.6. Relación entre el parámetro γ y el índice de fijación F_{st}

El índice de fijación F_{st} (Wright, 1943) es una medida de la diversidad de un conjunto de poblaciones con respecto a una de referencia que suele ser un conjunto o pool de todas las poblaciones. En este enfoque, cada población se asume como una muestra aleatoria de todas las posibles que podrían haber sido muestreadas de acuerdo al proceso evolutivo descripto por el F_{st}. Conceptualmente, este índice es un parámetro a estimar y no un estadístico calculado a partir de los datos. La definición usual de F_{st} para un locus bialélico en particular es:

$$F_{st} = \frac{\sigma_p^2}{\tilde{p}(1-\tilde{p})}, \qquad [2.28]$$

Donde σ_p^2 es la varianza de las frecuencias alélicas a través de las poblaciones y $\tilde{p}$ es la frecuencia alélica de la población combinada conceptual. Considerando que la varianza de las frecuencias alélicas se aplica entre loci y no entre poblaciones, se sigue que $\tilde{p} = 0,5$ debido a que los alelos de referencia son tomados al azar. En este caso:

$$F_{st} = \frac{\sigma_p^2}{\tilde{p}(1-\tilde{p})} = \frac{\sigma_p^2}{(0.5)^2} = 4\,\sigma_p^2 = \frac{\gamma}{2}. \qquad [2.29]$$

La interpretación de la relación existente entre F_{st} y γ viene de Jacquard (1974), quién llamó "coeficiente de consanguinidad de una población" al valor $\gamma/2$. Cockerham (1969) modeló $\gamma/2 = \theta_l = F_{st}$ como una correlación intraclase, "la coancestría de la línea consigo misma", es decir, la probabilidad que dos gametas tomadas al azar de una población sean idénticas entre sí. En consecuencia, es lógico considerar que la relación aditiva (dos veces la coancestría) de un grupo consigo mismo sea $\gamma = 2\theta_l = 8\sigma_p^2$. Esta es la interpretación del coeficiente $\gamma/2$ en Legarra *et al.* (2015). Nótese que el supuesto de $\tilde{p} = 0,5$ se cumple automáticamente si los alelos de referencia son tomados al azar entre los loci. Es decir, que no son los más frecuentes ni los menos observados.

Alternativamente, Legarra *et al.* (2015) mostraron que, para una población con una relación media igual a γ, la heterocigosidad promedio es $1-\gamma/2$. Es decir que la varianza se reduce $\gamma/2$ de la población conceptual con heterocigosidad 1. En consecuencia, $\gamma/2$ puede interpretarse como F_{st} si el mismo es tomado como una medida de la homocigosidad.

Capítulo 3

Estimación de los parámetros γ empleando información de marcadores moleculares y genealogía

Estimación de los parámetros γ empleando información de marcadores moleculares y genealogía

3.1. INTRODUCCIÓN

En este capítulo se aborda el problema de la estimación de γ y el cálculo de *s*, definidos en el precedente. Es necesario contar con buenas estimaciones de ambos parámetros para lograr una implementación simple y adecuada del modelo con MF en selección genómica. Ahora bien, la importancia de la estimación de γ no sólo se circunscribe a la predicción de los valores de cría incluyendo a los MF en el modelo, sino que el valor del parámetro posee importancia por sí mismo dada su relación con el índice de fijación F_{st}, como muestran los resultados teóricos del capítulo 2. Tal cual lo comentado en el capítulo precedente, el hecho de contar con valores para γ permite caracterizar la variabilidad genética de la población motivo por el cual es de interés contar con estimaciones precisas del parámetro. Por su parte, *s* es una medida global de la heterocigosis, y es de importancia en el cálculo de la matriz de estructura de covarianzas al incluir uno o varios MF en selección genómica.

Al momento de introducir γ, Christensen (2012) propuso, estimar el parámetro por máxima verosimilitud, con la ayuda de la información genómica. El algoritmo requiere métodos de búsqueda numérica como los de Monte Carlo. Como alternativa, Legarra *et al.* (2015) propusieron un método de momentos (MM) con estadísticos resumen de las matrices de relaciones genómicas y de pedigrí. La sencillez de dicho método radica en que sólo requiere medidas resumen de las mencionadas matrices sin necesidad de construirlas. En consecuencia, el impacto en el costo de cálculo es considerable. Como la citada metodología sólo permite obtener aproximaciones al parámetro, en este capítulo se proponen alternativas de estimación, sobre la base de los resultados teóricos del capítulo 2, en donde se establece la relación entre γ y la varianza de las frecuencias alélicas base. Las propuestas de este capítulo pretenden abordar dos casos que pueden presentarse a la hora de trabajar con MF en las evaluaciones genómicas: i) contar con una única población y, en consecuencia, involucra a un solo MF; ii) contar con múltiples poblaciones, caso en el que es necesario incluir múltiples MF. Luego, en una segunda instancia, se evalúa el comportamiento de dichos métodos por medio de una simulación de datos de una población bovina lechera bajo selección. Se compara la performance de los métodos entre ellos y con la metodología propuesta por Legarra *et al.* (2015), MM, con el objetivo de determinar cuál resulta la mejor alternativa a la hora de estimar γ a partir de la información genómica y, en algunos casos, combinándola con la genealógica. Determinar qué método presenta el mejor comportamiento permite establecer la metodología a emplear al momento de predecir los EBV incluyendo los MF en el modelo y hacer recomendaciones al respecto. Este tema será abordado detalladamente más adelante.

3.2. MÉTODOS

3.2.1. Estimación de γ para una única población

En esta sección se presentan métodos alternativos al MM presentado por Legarra *et al.* (2015) para la estimación del parámetro γ. Los detalles de dicho método fueron presentados en la sección 2.2.3, del capítulo 2. Aquí, se presentan nuevas propuestas y posteriormente se procede a la evaluación y comparación de la performance en términos predictivos, tanto de los métodos aquí propuestos como del MM.

El parámetro γ es proporcional a la varianza de las frecuencias alélicas en la población base. En el caso de contar con información genómica para dicha población, computar las frecuencias alélicas y la estimación de γ sería trivial. Ahora bien, en casos reales no suele contarse con información genómica de los animales de la población base, motivo por el cual es necesario disponer de metodologías para estimar γ. A continuación, se proponen y describen métodos de estimación de dicho parámetro para los casos en que la única información genómica disponible es aquella de animales relacionados entre sí y de generaciones recientes, en muchos casos distantes de la población base.

3.2.1.1. Enfoque "ingenuo": asumiendo que no existe estructura de pedigrí.

El modelo más sencillo asume que los animales genotipados no se encuentran relacionados entre sí y constituyen la población base. Para el locus *i*, sea $\boldsymbol{m_i}$ un vector con los conteos de alelos de cada animal (del inglés, *gene content*) codificados como 0, 1 ó 2. Su valor esperado es $\mu_i = 2p_i$, donde p_i es la frecuencia alélica. Para cada locus, μ_i se estima como el promedio observado de $\boldsymbol{m_i}$ y luego se calcula la $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ como la varianza empírica empleando *n* loci de $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$. Dado que $p_i = \mu_i/2$, entonces

$$\mathrm{Var}(p_i) = \sigma^2_p = \tfrac{1}{4}\mathrm{Var}(\hat{\boldsymbol{\mu}}) \qquad [3.1]$$

Y el parámetro γ puede estimarse como

$$\hat{\gamma} = 8\,\sigma^2_p = 2\,\mathrm{Var}(\hat{\boldsymbol{\mu}}) \qquad [3.2]$$

3.2.1.2. Considerando la estructura de pedigrí.

En el locus *i*, el contenido génico puede considerarse como un carácter cuantitativo, la "media de $\boldsymbol{m_i}$ en la población base" igual a $2p_i$, donde p_i es la frecuencia alélica en la población base y la varianza genética es $2p_iq_i$ (McPeek *et al.*, 2004; Gengler *et al.*, 2007; Forneris *et al.*, 2015). Cockerham (1969) mostró que la covarianza del contenido génico del marcador *i* entre los individuos *j* y *k* es una función de su relación ($\boldsymbol{A}_{jk}$): $\mathrm{Cov}\left(m_{i,j}, m_{i,k}\right) = \boldsymbol{A}_{jk} 2p_iq_i$. En consecuencia, puede emplearse un modelo lineal del siguiente modo:

$$\boldsymbol{m}_i = \mathbf{1}\mu_i + \boldsymbol{W}\,\boldsymbol{u}_i + \boldsymbol{e} \qquad [3.3]$$

donde $\boldsymbol{W}$ es una matriz de incidencia que permite relacionar a los individuos del pedigrí con los genotipos observados y $\boldsymbol{u}_i$ es la desviación de cada individuo de la media μ_i para todos los individuos (McPeek *et al.*, 2004; Gengler *et al.*, 2007; Forneris *et al.*, 2015). Al asumir normalidad multivariada:

$$\mu \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{\mu}) \text{ y } \boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{A}(2p_iq_i)) = N(\boldsymbol{0}, \boldsymbol{A}\sigma^2_{m_i}) \,. \qquad [3.4]$$

De modo equivalente, para el conjunto de individuos genotipados (identificado como el grupo 2), $\boldsymbol{u}_{2,i} \sim N(\boldsymbol{0}, \boldsymbol{A}_{22}(2p_iq_i))$ donde $\boldsymbol{A}_{22} = \boldsymbol{WAW}'$ es la matriz de relaciones aditivas que incluye sólo a los animales genotipados. A partir de esta formulación existen dos estrategias posibles para estimar σ^2_{μ}. La primera es por mínimos cuadrados generalizados (GLS) y la segunda por máxima verosimilitud (ML). A continuación, se describen ambos:

3.2.1.2.1. Mínimos cuadrados generalizados (GLS)

Este enfoque ignora la distribución a priori de $\boldsymbol{\mu}$ y permite estimar cada μ_i bajo la especificación de un "efecto fijo" mediante un estimador lineal insesgado de mínima varianza (BLUE o, equivalentemente, GLS) de μ_i para cada uno de los loci separadamente. Una posibilidad es emplear $\boldsymbol{A}^{-1}$ abarcando toda la genealogía y las MME (McPeek *et al.*, 2004; Gengler *et al.*, 2007; Forneris *et al.*, 2015). De modo equivalente, la expresión de GLS es:

$$\hat{\mu}_{\mathrm{i}} = \left(\boldsymbol{1}'\boldsymbol{A}_{22}^{-1}\boldsymbol{1}\,\sigma^{-2}_{\mathrm{m_i}}\right)^{-1}\boldsymbol{1}'\boldsymbol{A}_{22}^{-1}\boldsymbol{m}_{\mathrm{i}}\,\sigma^{-2}_{\mathrm{m_i}} = \left(\boldsymbol{1}'\boldsymbol{A}_{22}^{-1}\boldsymbol{1}\right)^{-1}\boldsymbol{1}'\boldsymbol{A}_{22}^{-1}\boldsymbol{m}_{\mathrm{i}} \qquad [3.5]$$

donde $\left(\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\right)$ es la suma de los elementos de $\mathbf{A}_{22}^{-1}$, $\sigma_{m_i}^2 = 2p_iq_i$ y $\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{m}_i$ es una suma ponderada de los genotipos. Entonces, σ_μ^2 es estimada como $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ porque $p_i = \mu_i/2$, $\sigma_p^2 = \sigma_\mu^2/4$ y, consecuentemente, $\hat{\gamma} = 2\hat{\sigma}_\mu^2$.

3.2.1.2.2. Máxima verosimilitud (ML)

Si las frecuencias alélicas en la población base poseen una distribución, μ_i puede ser considerada como muestreada de una distribución normal, $\boldsymbol{\mu} \sim N(\mathbf{0}, \mathbf{I}\sigma_\mu^2)$. De este modo σ_μ^2 es un componente de varianza que puede ser estimado por máxima verosimilitud. Las ecuaciones para valores dados de σ_μ^2 y $\sigma_{m_i}^2 = 2p_iq_i$ son

$$\left(\mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{1}\,\sigma_{m_i}^{-2} + \sigma_\mu^{-2}\right)^{-1}\hat{\mu}_i = \mathbf{1}'\mathbf{A}_{22}^{-1}\mathbf{m}_i\,\sigma_{m_i}^{-2} \qquad [3.6]$$

Para obtener ML es posible implementar un algoritmo EM (del inglés *expectation-maximization*) (Mäntysaari y Van Vleck, 1989). A partir de valores iniciales para σ_μ^2 y $\sigma_{m_i}^2$, se iteran los siguientes pasos hasta que se halle convergencia:

1) Para cada marcador *i*,

a) Estimar $\hat{\mu}_i = (\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1} + \sigma_\mu^{-2})^{-1}\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{m}_i$,

b) Almacenar $PEV_i(\hat{\mu}_i) = (\mathbf{1}'\mathbf{A}_{22}^{-1}\sigma_{m_i}^{-2}\mathbf{1} + \sigma_\mu^{-2})^{-1}$

c) Actualizar $\sigma_{m_i}^2$ como $\sigma_{m_i}^2 = 2\hat{p}_i\hat{q}_i$ con $\hat{p}_i = \hat{\mu}_i/2$;

2) Actualizar σ_μ^2 como $\hat{\sigma}_\mu^2 = \frac{1}{n}\left(\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + \sum PEV_i(\hat{\mu}_i)\right)$. La segunda parte de la expresión corresponde a la traza $\mathrm{tr}(\mathbf{IC})$. $\mathbf{I}$ corresponde a la matriz de relaciones a través de los niveles de $\boldsymbol{\mu}$ y $\mathbf{C}$ es la matriz de covarianza del error de predicción de $\hat{\boldsymbol{\mu}}$. Dado que sólo se requieren los elementos diagonales de $\mathbf{C}$ en la traza $\mathrm{tr}(\mathbf{IC})$, sus elementos $PEV_i(\hat{\mu}_i)$ pueden calcularse separadamente para cada locus.

Una vez alcanzada la convergencia y con las estimaciones obtenidas para $\hat{\sigma}^2_{\mu}$ es posible estimar γ según $\hat{\gamma} = 2\,\hat{\sigma}^2_{\mu}$. Esto permite obtener la misma estimación que con el método basado en una función de verosimilitud Wishart (Christensen, 2012) con $s = n/2$. Dicha igualdad fue verificada empleando los datos simulados y los resultados se presentan en la sección 3.4.

3.2.2. Estimación de Γ para múltiples poblaciones

Si se consideran *t* poblaciones base, el componente de varianza σ^2_{μ} se generaliza a $\boldsymbol{\Sigma}_0$, una matriz de orden $t \times t$ de varianzas y covarianzas entre medias $\mu_i^{[b]}$ para el marcador *i* en la población *b*. Entre poblaciones,

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \sigma^2_{\mu^{[1]}\mu^{[1]}} & \sigma_{\mu^{[1]}\mu^{[2]}} & \cdots & \sigma_{\mu^{[1]}\mu^{[t]}} \\ & \sigma^2_{\mu^{[2]}\mu^{[2]}} & \cdots & \sigma_{\mu^{[2]}\mu^{[t]}} \\ & & \ddots & \vdots \\ & \text{simétrica} & & \sigma^2_{\mu^{[t]}\mu^{[t]}} \end{bmatrix} \quad \text{y} \quad \hat{\boldsymbol{\Gamma}} = 2\hat{\boldsymbol{\Sigma}}_0 . \tag{3.7}$$

3.2.2.1. Enfoque ingenuo: asumiendo que no existe estructura de pedigrí.

Si las relaciones entre los individuos son ignoradas:

$$\boldsymbol{m}_i = \boldsymbol{Q}\boldsymbol{\mu}_i + \boldsymbol{e}_i , \tag{3.8}$$

Las filas de la matriz $\boldsymbol{Q}$ suman a uno para poder asignar cada uno de los individuos a "fracciones" de poblaciones. El vector $\boldsymbol{\mu}_i$ contiene *t* elementos que consisten en los promedios de cada población. Para cada locus, $\boldsymbol{\mu}_i$ puede estimarse empleando mínimos cuadrados y la matriz de covarianzas de $\boldsymbol{\mu}_i$ a través de los loci permite obtener una estimación de $\boldsymbol{\Sigma}_0$. Por ejemplo, para dos poblaciones $\hat{\boldsymbol{\Sigma}}_0 = \text{Cov}(\boldsymbol{\mu}^{[1]}, \boldsymbol{\mu}^{[2]})$ corresponde a una matriz de orden 2×2.

3.2.2.2. Considerando la estructura de pedigrí

Si no existen cruzas entre individuos de poblaciones diferentes en la genealogía, la estimación de la varianza de las frecuencias alélicas en cada población base (*b*) puede realizarse independientemente del siguiente modo:

$$\boldsymbol{m}_i^b = \mathbf{1}\,\mu_i^{[b]} + \boldsymbol{W}^b\,\boldsymbol{u}_i^b + \boldsymbol{e} \qquad [3.9]$$

con $\boldsymbol{u}_i^b \sim N\left(\boldsymbol{0}, \boldsymbol{A}^b\left(2p_i\left(1-p_i\right)\right)\right)$ donde $\boldsymbol{A}^b$ es la matriz de relaciones calculadas empleando la información de pedigrí entre los individuos de la población *b*, y el análisis continua como en el caso de una única población, tal como se presentó en la sección 3.2.1. Luego, $\hat{\boldsymbol{\Sigma}}_0$ es estimada como la matriz observada de covarianzas para $\hat{\mu}_i^b$ para todos los loci.

Ahora bien, si existen animales cruza de dos poblaciones, existen dos alternativas de estimación que se detallan a continuación.

3.2.2.2.1. Mínimos cuadrados generalizados (GLS)

Por un lado, la primera alternativa (Forneris *et al.*, 2015) es emplear un modelo con grupos genéticos (Thompson, 1979; Quaas, 1988), como $\boldsymbol{m}_i = \boldsymbol{Q}\mu_i + \boldsymbol{W}\boldsymbol{u}_i + \boldsymbol{e}$ donde $\boldsymbol{Q}_{k,b}$ contiene la fracción de la ancestría *b* en el individuo *k*. Esto ignora el hecho de que la variancia del contenido génico ($2p_iq_i$) es heterogénea entre razas y cruzas. La segunda alternativa es más exacta. En este caso se emplea la representación en la que los valores de cría se descomponen en dos componentes, uno intrarracial y otro interracial, del siguiente modo:

$$\boldsymbol{m}_i = \boldsymbol{Q}\mu_i + \sum_b \boldsymbol{W}^b\boldsymbol{u}_i^b + \sum_{b,b',b>b'} \boldsymbol{W}^{b,b'}\boldsymbol{u}_{\mathrm{i}}^{b,b'} + \boldsymbol{e}\,, \qquad [3.10]$$

A tal fin se emplean matrices de relaciones parciales para los vectores $\boldsymbol{u}_{\mathrm{i}}^b$ y $\boldsymbol{u}_{\mathrm{i}}^{b,b'}$. Pueden obtenerse los BLUE (estimador lineal insesgado de mínima varianza, del inglés *Best linear unbiased estimator*) de $\boldsymbol{\mu}_{\mathrm{i}}$ y luego estimarse $\hat{\boldsymbol{\Sigma}}_0$, tal como se describió anteriormente.

3.2.2.2.2. Máxima Verosimilitud (ML)

Análogo al caso de una única población, se puede obtener una estimación actualizada empleando el algoritmo EM utilizando formulaciones multicarácter (Mäntysaari y Van Vleck, 1989), donde PEC es la (co)varianza del error de predicción. Por ejemplo, para el caso de dos poblaciones tenemos:

$$\hat{\boldsymbol{\Sigma}}_0 = \begin{bmatrix} \boldsymbol{\mu}^{[1]\prime}\boldsymbol{\mu}^{[1]} & \boldsymbol{\mu}^{[1]\prime}\boldsymbol{\mu}^{[2]} \\ \boldsymbol{\mu}^{[2]\prime}\boldsymbol{\mu}^{[1]} & \boldsymbol{\mu}^{[2]\prime}\boldsymbol{\mu}^{[2]} \end{bmatrix}. \qquad [3.11]$$

Para implementar este enfoque es necesario seguir los siguientes pasos:

1) Para cada marcador *i*:

a) Estimar $\hat{\boldsymbol{\mu}}_i = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{Q}\sigma_{m_i}^{-2}\right)^{-1}\boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\boldsymbol{m}_i\sigma_{m_i}^{-2}$,

b) Almacenar $PEC_i(\hat{\boldsymbol{\mu}}_i) = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{Q}'\boldsymbol{A}_{22}^{-1}\sigma_{m_i}^{-2}\boldsymbol{Q}\right)^{-1}$,

c) Actualizar $\sigma_{m_i}^2$ como $\hat{\sigma}_{m_i}^2 = 2\hat{p}_i^*\left(1-\hat{p}_i^*\right)$ con $\hat{p}_i^* = \frac{1}{Nb}\sum_{b=1,Nb}\frac{\hat{\mu}_i^b}{2}$;

2) Actualizar $\boldsymbol{\Sigma}_0$ empleando productos cruzados entre y dentro de las poblaciones involucradas. En el caso de dos poblaciones, tendríamos:

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\left(\begin{bmatrix} \hat{\boldsymbol{\mu}}^{[1]\prime}\hat{\boldsymbol{\mu}}^{[1]} & \hat{\boldsymbol{\mu}}^{[1]\prime}\hat{\boldsymbol{\mu}}^{[2]} \\ \hat{\boldsymbol{\mu}}^{[2]\prime}\hat{\boldsymbol{\mu}}^{[1]} & \hat{\boldsymbol{\mu}}^{[2]\prime}\hat{\boldsymbol{\mu}}^{[2]} \end{bmatrix} + \sum_{i=1,n} PEC_i\right). \qquad [3.12]$$

El paso 1) involucra una aproximación en el apartado c) al asumir que $\sigma_{m_i}^2 = 2p_iq_i$ es la misma para todas las poblaciones base. Lo mismo ocurre con GLS presentado anteriormente y puede mejorarse empleando matrices de relaciones parciales. Se planea abordar esta propuesta de mejora en futuros trabajos de investigación.

3.2.3. Evaluación de los métodos para estimar γ

El parámetro γ de una población simulada se estimó empleando cada uno de los cuatro métodos presentados anteriormente. Por un lado, se empleó el método "ingenuo" que no contempla ninguna estructura de pedigrí y por otro lado dicha estructura fue considerada por los otros tres métodos: MM (propuesta por Legarra *et al.* 2015 y descripta en el capítulo 2 de esta tesis), ML y GLS. Las estimaciones obtenidas de γ por cada método se compararon entre sí y con el valor real de la base de datos simulada que se describe en la próxima sección.

3.3. MATERIALES

3.3.1 Base de datos simulada

Se realizó una simulación estocástica empleado el programa QMSim (Sargolzaei y Schenkel, 2009) con el objetivo de evaluar la calidad de las estimaciones que se obtienen al emplear los métodos descriptos en la sección 3.2. Se simuló un rodeo de bovinos de leche bajo selección, de manera similar a la efectuada por Vitezica *et al.* (2011). Tal como se presenta en la Figura 3.1, se simuló una población en dos etapas. Primero se simularon generaciones históricas bajo deriva y mutación para crear un nivel de LD comparable al real y, en una segunda etapa, se generó la población reciente bajo selección.

Primero, se generaron 100 generaciones de una población histórica con un tamaño efectivo de 100 durante las primeras 95 generaciones, seguidas de una expansión gradual durante las últimas cinco generaciones hasta alcanzar un tamaño efectivo de 3000. En lo que respecta al genoma, se simularon 30 cromosomas de 100 cM y 40.000 marcadores bialélicos distribuidos al azar a lo largo de los cromosomas en la primera generación de la población histórica. Los 40.000 marcadores fueron muestreados de un conjunto mayor de 90.000 marcadores. El objetivo de este muestreo fue obtener frecuencias alélicas de una distribución Beta (2,2), similar a aquellas observadas en bases de datos reales de bovinos lecheros, buscando que el parámetro γ tuviese un valor verdadero cercano a 0,40. El carácter simulado (producción de leche) se encontraba afectado por 1500 QTLs y la heredabilidad del mismo fue 0,30. Los efectos de los alelos de los QTLs se muestrearon de una distribución Gamma, cuyo parámetro de forma fue 0,40. Se aplicó una tasa de mutación de $2{,}5 \times 10^{-5}$ por locus por generación (Solberg *et al.*, 2008) para los marcadores y los QTLs, asumiendo un modelo de mutación recurrente.

Posteriormente, se simularon 10 generaciones recientes superpuestas bajo selección. En cada una se aparearon 200 machos con 2600 hembras de modo de producir 2600 crías en apareamientos dirigidos: las hembras con mayores valores de cría predichos (EBV) fueron cruzadas con los machos de mayores EBV. Los animales se seleccionaron en función de su valor de cría predicho empleando BLUP con información genealógica, tal

como se realiza en las evaluaciones genéticas tradicionales (PBLUP). En cada generación de la población reciente se reemplazó el 40% de los machos y el 20% de las hembras reteniendo animales jóvenes selectos. No se impuso ninguna restricción para minimizar la consanguinidad, con lo cual se observaron individuos con un alto coeficiente de consanguinidad (*F*). Un total de 100 individuos mostraron un $F > 0{,}20$, principalmente entre los animales de la última generación, y algunos superaron 0,40. Los verdaderos valores de cría (TBV) y la genealogía (con un total de 28.800 animales) estaban disponibles para todos los animales de las 10 generaciones de la población reciente. Todas las hembras contaron con registros fenotípicos, exceptuando aquellas de la última generación, lo que dio como resultado un total de 14.300 de animales con datos. Los 840 padres de las hembras con datos fenotípicos se encontraban genotipados, así como también los 2600 individuos de la generación nueve (con datos fenotípicos) y 2600 en la generación 10 (sin registros fenotípicos). En total, 6038 animales contaban con información genómica. Se generaron 20 réplicas independientes de la simulación. Las principales características de la población simulada se detallan en la Figura 3.1.

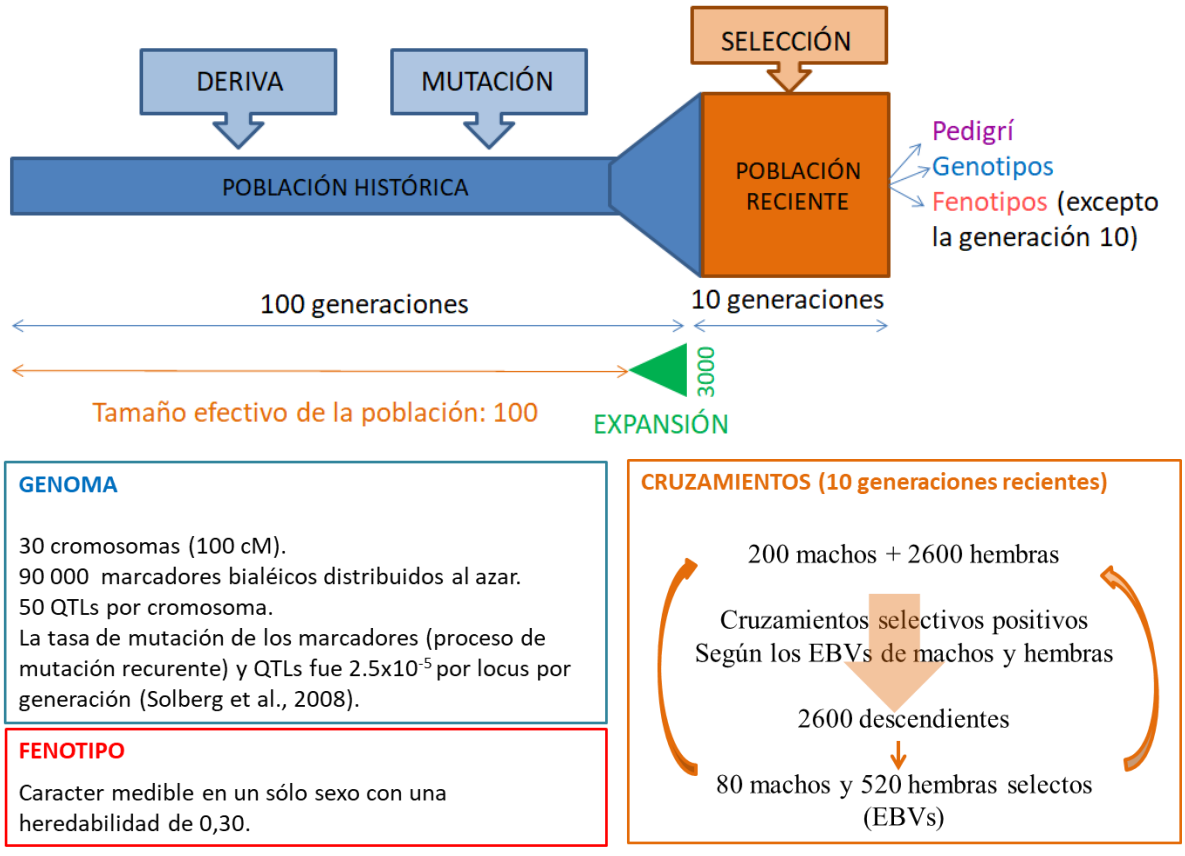


Figura 3.1. Esquema de la estructura poblacional de la base de datos simulada y resumen de las principales características del genoma, fenotipo y esquema de cruzamientos.

Esta base de datos simulada se empleó en un análisis de dos partes y con dos objetivos diferentes. En una primera etapa, se compararon diferentes métodos para estimar γ. Los resultados de estos análisis se presentan a continuación en este capítulo. Luego, en una segunda etapa, se evaluó la calidad de las predicciones genómicas de los animales de la décima generación (candidatos a la selección) al incluir un MF en el modelo. Dicha evaluación se realizó en términos de exactitud, sesgo e "inflación" de las predicciones. En el capítulo 4 se describe el objetivo de este segundo análisis y los resultados obtenidos en mayor detalle.

3.4. RESULTADOS

La Figura 3.2. muestra los diagramas de la diferencia entre el valor del parámetro γ estimado y el verdadero (conocido por simulación) para cada uno de los cuatro métodos de estimación (ingenuo, MM, ML y GLS) considerando las 20 réplicas. Recordemos que la simulación fue pensada de manera tal que γ = 0,40. Tal como se desprende de la Figura 3.2., tanto ML como GLS estimaron el parámetro con gran exactitud. El método MM subestimó γ, mientras que el método ingenuo lo sobreestimó.

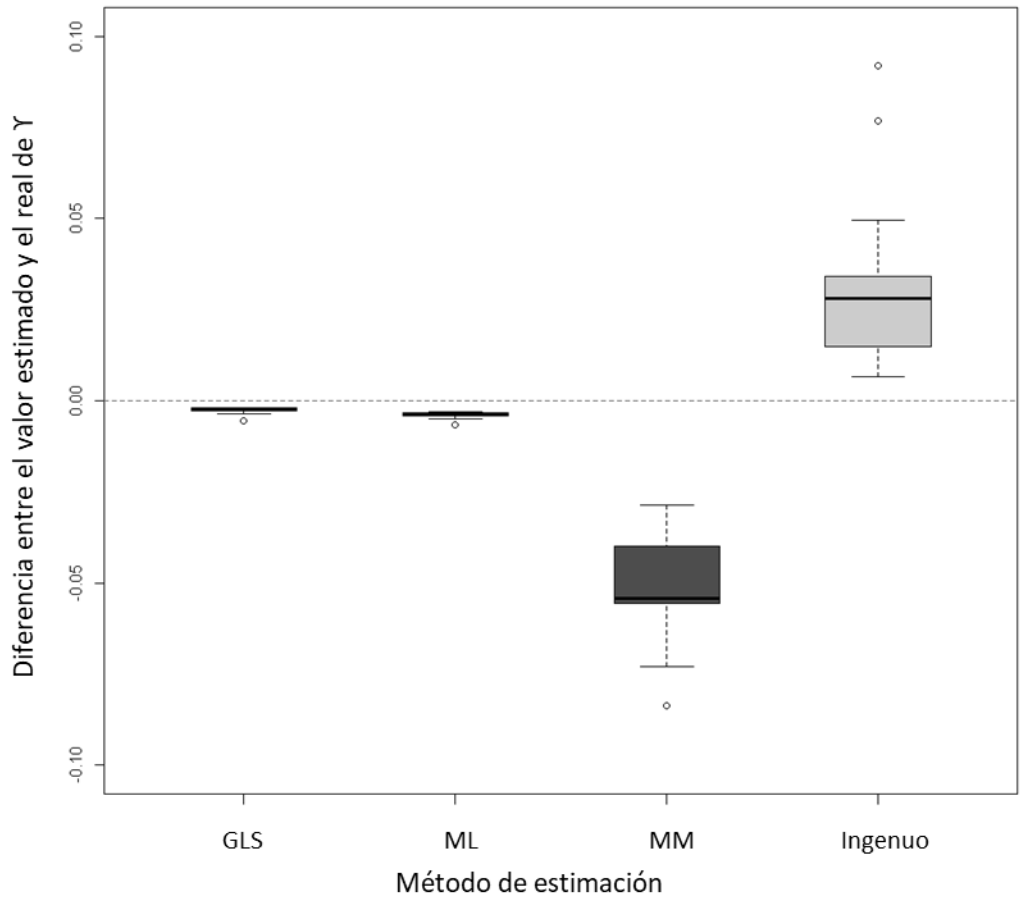


Figura 3.2. Diferencias entre el valor estimado de γ y su valor verdadero para las 20 réplicas de la simulación. El parámetro γ se estimó empleando cuatro métodos: mínimos cuadrados generalizados (GLS), máxima verosimilitud (ML), método de momentos (MM) y el método ingenuo.

La Figura 3.3. presenta la curva del logarítmo de la verosimilitud Wishart obtenida para una de las réplicas de la simulación al emplear el método propuesto por Christensen (2012) con $s = n/2$, siendo *n* el número de marcadores, para estimar γ. Se presentan los resultados para sólo una de las réplicas de la simulación a modo de ejemplo ya que en todos los casos la tendencia fue similar. Nótese que el valor máximo de la verosimilitud se observa para valores cercanos a $\gamma = 0{,}40$ (valor real del parámetro), tal como ocurre con el ML. De este modo se muestra, como se mencionara en la sección 3.2.1.2.2, que ML permite obtener estimaciones que se corresponden con las generadas por el método de Christensen (2012).

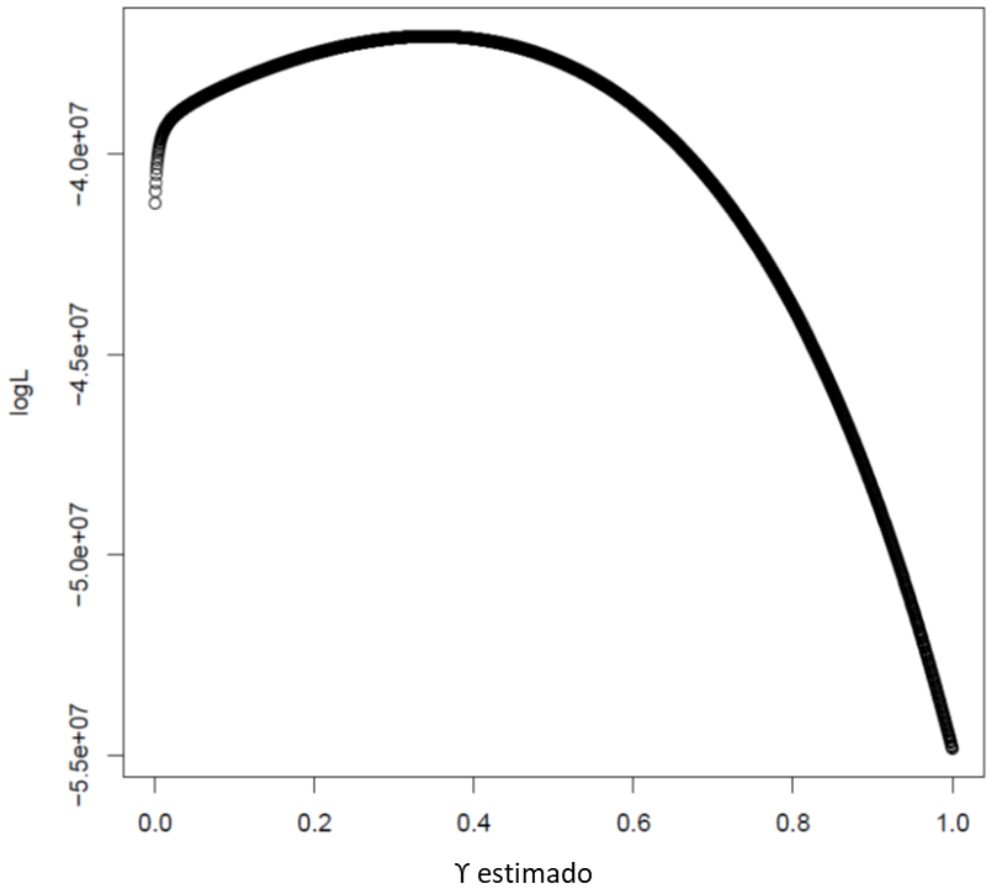


Figura 3.3. Estimación de γ empleando el método basado en una función de verosimilitud Wishart (Christensen, 2012). Se consideró $s = n/2$, donde *n* es el número de marcadores.

3.5. DISCUSIÓN

Los resultados teóricos presentados en el capítulo 2 permiten proponer métodos alternativos para la estimación de las relaciones ancestrales de los individuos base, γ (para una única población) y **Γ** (para más de una población). Las propuestas se dan en el marco de los modelos lineales clásicos de la genética cuantitativa (Cockerham, 1969), tal como se ha empleado recientemente para el conteo de alelos (*gene content*) (Makgahlela et al, 2014; McPeek *et al.*, 2004; Gengler *et al.*, 2007; Forneris *et al.*, 2015). Este enfoque es sencillo de comprender y fácil de implementar. Además, **Γ** puede interpretarse, tal como sucede con la heredabilidad, como un parámetro no observable de la población base que no cambia al

considerar más información, a diferencia de la estimación que si se puede modificar. En consecuencia, se puede estimar precisamente $\boldsymbol{\Gamma}$ y emplearse reiteradamente sin necesidad de tener que volver a estimarse, tal como suele llevarse a cabo en las evaluaciones genéticas ganaderas con los componentes de varianza. Esto contrasta directamente con el "centrado" comúnmente efectuado al momento de calcular la matriz $\boldsymbol{G}$ de relaciones genómicas, que cambia con la introducción de cada nuevo genotipo. Ahora bien, si se conocieran todas las frecuencias alélicas de la base no sería necesario emplear MF ya que las matrices genómicas podrían construirse apropiadamente, sin incertidumbre alguna (Makgahlela *et al.*, 2014).

Entre los distintos métodos para estimar γ, la metodología "ingenua" y MM produjeron estimaciones sesgadas del parámetro. El primero sobreestimó el parámetro en todos los casos, mientras que el segundo lo subestimó. En relación con el método ingenuo, el sesgo se origina al ignorar la estructura poblacional representada por el pedigrí. Este método no considera que las frecuencias alélicas pueden tomar valores extremos a medida que avanzan las generaciones, por causa de la deriva génica. Por otra parte, si bien el MM emplea información genealógica, asume implícitamente que los individuos genotipados son una muestra aleatoria de una generación en particular. En este sentido, y tal como demuestran los resultados, los métodos que permiten estimar el parámetro considerando la estructura poblacional (genealogía) son los que mejor se comportan (ML y GLS).

Capítulo 4

Impacto del empleo de los metafundadores en selección genómica

Impacto del empleo de los metafundadores en la selección genómica

4.1. INTRODUCCIÓN

En este capítulo se evalúa el impacto del empleo de los MF en selección genómica. Como se mencionó en el capítulo 2, Legarra *et al.* (2015) propusieron el empleo de MF como una alternativa viable para resolver problemas frecuentemente observados a la hora de implementar la selección genómica. Entre estos aspectos problemáticos se destaca la falta de compatibilidad entre las matrices de relaciones genómicas y las de pedigrí. Fue en este contexto en el que surgió el enfoque de los MF como una solución a estas incompatibilidades, derivada de los trabajos de Jacquard (1969, 1974), VanRaden (1992), Aguilar y Misztal (2008), VanRaden *et al.* (2011), Colleau y Sargolzaei (2011) y Christensen (2012).

Ahora bien, la originalidad de esta investigación reside en evaluar el desempeño del modelo incorporando MF en términos predictivos y de estimación de componentes de varianza, empleando para ello un conjunto de datos simulados de una población bovina lechera bajo selección.

4.2. MATERIALES Y MÉTODOS

4.2.1. Base de datos simulada

Para evaluar el desempeño del método se empleó la base de datos simulada descripta en el capítulo 3 de esta tesis. A continuación se resumen brevemente sus principales características. La población simulada contaba con un total de 28800 individuos en la genealogía, 6038 de los cuales disponían de información genómica. Se simularon 100 generaciones históricas y 10 recientes. Sólo las hembras contaron con registros fenotípicos. Los machos y los animales de las dos últimas generaciones poseían información genómica. El genotipo de cada animal contaba con 40.000 marcadores SNP. La décima generación de la población reciente no contaba con registros fenotípicos y fue utilizada para evaluar el desempeño de los métodos en términos de exactitud, sesgo e inflación de las predicciones. Se llevaron a cabo 20 réplicas de la simulación y en cada una de ellas se evaluaron cada uno de dichos aspectos. La descripción detallada de la simulación se encuentra en el capítulo 3, sección 3.3. A continuación se presenta un resumen de sus principales características dentro del Cuadro 4.1.

Cuadro 4.1. Resumen de la estructura poblacional de la base de datos simulada y de las principales características del genoma, fenotipo y esquema de apareamientos.

Objetivo: simular un rodeo de bovinos de leche bajo selección

Nº réplicas: 20

Población

1. Generaciones históricas

Nº generac. totales: 100

Nº generac. etapa 1: 95 (N_e =100)

Nº generac. etapa 2: 5 (expansión a N_e =3000)

Procesos: deriva y mutación

2. Generaciones recientes

Nº generaciones totales: 10

Nº hembras fundadoras: 2600

Nº machos fundadores: 200

Nº hembras selectas por generación: 520

Nº machos selectos por generación: 80

Apareamientos dirigidos según EBV

Información disponible: genealógica, genotipos, fenotipos (excepto la generac. 10)

Genoma

Nº cromosomas: 30

Largo de cada cromosoma: 100 cM

Nº de QTLs por cromosoma: 50

Nº total SNPs: 40.000 (al azar en el genoma)

Tasa de mutación QTLs y SNPs: 2,5 x 10^{-5}

Fenotipo

Medible sólo en hembras

Heredabilidad: 0,30

4.2.2. Métodos de predicción genómicos

Los EVB de los candidatos a la selección en la generación 10 (con información genotípica pero sin datos fenotípicos) fueron estimados empleando cuatro métodos. El primero fue BLUP basado en la genealogía (PBLUP) tomando en cuenta los datos fenotípicos y los de pedigrí. El segundo método empleado fue ssGBLUP, en el que también se incluye la información genómica en la predicción. En este caso se empleó la corrección de Christensen *et al.* (2012) con el objetivo de igualar el promedio de los coeficientes de consanguinidad y relaciones genómicas y de pedigrí, el método usado por defecto en muchas de las aplicaciones prácticas (Christensen *et al.*, 2012; Masuda *et al.*, 2016). Cabe destacar que la implementación empleada no considera la consanguinidad al momento de calcular $\boldsymbol{A}^{-1}$ (Mehrabani-Yeganeh *et al.*, 2000), pero si lo hace con el cálculo de $\boldsymbol{A}_{22}^{-1}$ (ver más abajo la sección 4.2.3 para más detalles acerca del uso de estas matrices). El tercer método empleado fue ssGBLUP que incluye consanguinidad en el cómputo de $\boldsymbol{A}^{-1}$ y $\boldsymbol{A}_{22}^{-1}$ (ssGBLUP_F). El cuarto método fue ssGBLUP con un MF (ssGBLUP_M), empleando γ estimada por GLS, dado que resultó ser un método exacto al estimar dicho parámetro, tal como fuera demostrado en el capítulo 3.

Los cuatro métodos emplearon las siguientes matrices de parentesco inversas:

I. **PBLUP**:

$$\boldsymbol{A}^{-1} \qquad [4.1]$$

II. **ssGBLUP:**

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_a^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \qquad [4.2]$$

La matriz $\boldsymbol{G}_a$ fue calculada como sugirieron Christensen *et al.* (2012) y $\boldsymbol{A}^{-1}$ es construida ignorando la consanguinidad (Mehrabani-Yeganeh *et al.*, 2000).

III. **ssGBLUP_F:**

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_a^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \qquad [4.3]$$

La matriz $\boldsymbol{A}^{-1}$ fue calculada considerando la consanguinidad.

IV. **ssGBLUP_M:**

$$\boldsymbol{H}^{\gamma-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{\gamma-1} \end{bmatrix} \qquad [4.4]$$

siendo $\boldsymbol{G} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})' / s$ y $s = n/2$ (capítulo 2) y $\boldsymbol{A}^{\gamma}$ calculada como en [2.4].

Se brindan más detalles en el apéndice. Para el cómputo se empleó *blupf90* (Misztal *et al.*, 2002). En el caso concreto de ssGBLUP_M, $\boldsymbol{H}^{\gamma-1}$ se construyó externamente con software propio y luego se ingresó como archivo de entrada a *blupf90*. En la siguiente sección se describe con mayor grade de detalle el modo en que fueron calculadas todas las matrices descriptas en este apartado.

4.2.3. Cálculo de las matrices *H*

Para ssGBLUP y ssGBLUP_F, la matriz $\boldsymbol{H}^{-1}$ es construida del siguiente modo (Legarra *et al.*, 2009; Christensen y Lund, 2010):

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}_a^{*-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix}, \qquad [4.5]$$

con

$$\boldsymbol{G}_a^{*} = 0.95\,\boldsymbol{G}_a + 0.05\boldsymbol{A}_{22} = 0.95(a\,\boldsymbol{J} + b\,\boldsymbol{G}) + 0.05\,\boldsymbol{A}_{22} \qquad [4.6]$$

y $\boldsymbol{G} = \dfrac{(\boldsymbol{M} - \boldsymbol{P})(\boldsymbol{M} - \boldsymbol{P})'}{2\sum p_i q_i}$, tal como en VanRaden (2008), $\boldsymbol{M}$ contiene los genotipos codificados como {0,1,2} y los elementos de $\boldsymbol{P}$ corresponden a dos veces las frecuencias alélicas p_i. Estas son calculadas a partir de los genotipos observados de modo tal que $2p_i$ es igual a la media de la *i-ésima* columna de $\boldsymbol{M}$. Las constantes *a* y *b* toman valores tales que los promedios de los todos los elementos y los diagonales de las matrices $\boldsymbol{G}_a$ y $\boldsymbol{A}_{22}$ sean iguales (Christensen *et al.*, 2012) con el objetivo de hacerlas compatibles entre sí. El empleo de los ponderadores 0,95 y 0,05 permiten asegurar la inversión de $\boldsymbol{G}_a$. Los

elementos diagonales de la matriz $\boldsymbol{A}^{-1}$ deben calcularse empleando las siguientes contribuciones (Meuwissen y Luo, 1992):

i) Ningún padre conocido: 1

ii) Un padre conocido: $\left(0,75-\frac{F_{\text{conocido}}}{4}\right)^{-1}$

iii) Ambos padres conocidos: $\left(0,50-\frac{F_{\text{padre}}}{4}-\frac{F_{\text{madre}}}{4}\right)^{-1}$

O, de modo más compacto y general, $\left(0,50-\frac{F_{\text{padre}}}{4}-\frac{F_{\text{madre}}}{4}\right)^{-1}$ con $F_{\text{desconocido}}=-1$.

ssGBLUP emplea las opciones por default de la familia de programas *blupf90* (opción *random_type add_animal*). Emplea un método sencillo para crear $\boldsymbol{A}^{-1}$ ya que asume que, en todos los casos, los *F* en las expresiones anteriores toman valor cero (*F* = 0).

A diferencia de ssGBLUP, ssGBLUP_F emplea $\boldsymbol{H}^{-1}$ calculada del modo presentado al inicio de esta sección pero al construir $\boldsymbol{A}^{-1}$ (opción *random_type add_an_upg*) se considera la consanguinidad y, en consecuencia, se aplican correctamente las reglas presentadas.

ssGBLUP_M emplea la opción *random_type user_file* de la familia de programas *blupf90*. Esta opción permite considerar una matriz de relaciones calculada externamente que en este caso toma la siguiente forma:

$$\boldsymbol{H}^{\boldsymbol{\Gamma}-1}=k\left(\boldsymbol{A}^{\boldsymbol{\Gamma}-1}+\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}^{*-1}-\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-1}\end{bmatrix}\right) \qquad [4.7]$$

con $\boldsymbol{G}^{*}=0,95\boldsymbol{G}+0,05\,\boldsymbol{A}_{22}^{\Gamma}$ (con el objetivo de asegurar que $\boldsymbol{G}$ sea no singular), $\boldsymbol{G}=(\boldsymbol{M}-\boldsymbol{J})(\boldsymbol{M}-\boldsymbol{J})'/s$ y $s=n/2$, $\boldsymbol{M}$ contiene los genotipos codificados como {0,1,2}, $\boldsymbol{J}$ es una matriz cuyos elementos son unos, *n* es el número de marcadores, $\boldsymbol{A}^{\boldsymbol{\Gamma}-1}$ y $\boldsymbol{A}_{22}^{\boldsymbol{\Gamma}-1}$ se construyeron empleando programas propios sobre la base del trabajo de Legarra *et al.* (2015) y utilizando el valor estimado de γ por GLS. La consanguinidad es considerada en

ambas matrices ($\boldsymbol{A}^{\Gamma-1}$ y $\boldsymbol{A}_{22}^{\Gamma-1}$). La constante $k = 1 - \frac{\gamma}{2}$ permite colocar a la varianza genética asociada a los MF (escenario en el que asumen los animales base relacionados) en la misma escala que la varianza tradicional (asumiendo no relacionados a los animales base ya sea en $\boldsymbol{A}$ o $\boldsymbol{H}$).

4.2.4. Calidad de las predicciones genómicas

La calidad de las predicciones genómicas se evaluó para los 2600 candidatos a la selección de la generación 10 de la población simulada de bovinos lecheros bajo selección, ya descripta en la sección 3.3. En el Cuadro 4.1 se resumen las principales características de la base de datos simulada. La exactitud de cada método de predicción se cuantificó empleando el coeficiente de correlación de Pearson entre TBV y EBV. El sesgo de las predicciones se calculó como la diferencia en el promedio de los TBV y el de los EBV con respecto a la población base (en el caso de ssGBLUP_M, con respecto a la solución del MF o a cero para los tres métodos restantes). En consecuencia, el sesgo se relaciona al progreso genético estimado en los candidatos a la selección.

La inflación de la predicción, comúnmente conocida como sesgo, se calculó mediante el coeficiente de regresión de los TBV en los EBV. Estos dos estadísticos corresponden a los coeficientes b_0 y b_1 en la metodología de validación propuesta por Interbull (Mäntysaari *et al.*, 2010). La misma emplea la regresión

$$\text{TBV} = b_0 + b_1 \text{ EBV} + e \qquad [4.8]$$

Se dice que una predicción está "inflada" cuando $b_1 < 1$.Por otro lado, se calculó el error cuadrático medio (MSE) de predicción de los EBV como el cuadrado de la media de la diferencia entre TBV y EBV.

Cabe destacar que un método ideal debería maximizar la exactitud, presentar mínimo MSE, sesgo nulo y un coeficiente de regresión igual a uno (ausencia de inflación). Si bien son propiedades estadísticas deseables, también tienen impacto en la selección animal. Otro aspecto de gran importancia para el mejoramiento genético animal, principalmente a la hora de llevar a cabo la selección, es el ranking de los individuos según los valores de EBV de cada candidato a la selección. Es de vital importancia que el método de predicción asegure la elección de los mejores candidatos por su mérito genético y que este ranking no sea alterado por el método de predicción empleado. Es por este motivo que también se evaluaron los cambios en el ranking de los candidatos a la selección al variar el método de predicción. A tal efecto se empleó el coeficiente de correlación de Spearman entre métodos. Valores altos de dicha correlación sugieren que ambos métodos ordenan a

los animales de modo muy similar, mientras que valores bajos muestran alteraciones importantes del ranking.

Por otro lado, se evaluó el efecto de emplear diferentes valores de γ en las predicciones genómicas resultantes. El objetivo fue cuantificar su impacto en las predicciones de los valores de cría.

4.3. RESULTADOS

4.3.1. Calidad de las predicciones genómicas

En el Cuadro 4.2 y Figura 4.1a se presentan las correlaciones entre TBV y EBV de los individuos candidatos a la selección (generación 10) para cada método de predicción. Comparado con PBLUP, las metodologías ssGBLUP_F y ssGBLUP_M permitieron alcanzar un incremento en la exactitud de aproximadamente 23 puntos absolutos. Esto muestra una mejora considerable en la predicción al incluir información molecular, y la posibilidad de generar una ganancia pequeña adicional al incluir al MF. ssGBLUP produjo una pequeña pérdida en exactitud comparado con los niveles alcanzados por ssGBLUP_F y ssGBLUP_M.

Por otro lado, el Cuadro 4.2 y la Figura 4.1b muestran los coeficientes de regresión de los TBV en EBV, que permiten medir el grado de inflación de las predicciones de cada método (será mejor aquel método que permita alcanzar valores muy cercanos a 1). Tanto PBLUP como ssGBLUP_F mostraron valores de la regresión cercanos a la unidad, es decir, las predicciones obtenidas con ambos métodos fueron las menos “infladas”. Al incluir información genómica en la predicción (ssGBLUP), se observaron coeficientes de regresión menores a la unidad (predicciones sobreestimadas), pero al incluir el MF (ssGBLUP_M) se obtuvieron valores más cercanos a uno. Los métodos ssGBLUP_M y ssGBLUP_F presentaron valores menores de desvío estándar (DS) para la inflación comparados con los otros dos métodos. De estos dos últimos, ssGBLUP fue el que presentó mayor variabilidad.

Cuadro 4.2 Exactitud (correlación entre TBV y EBV), inflación (coeficiente de regresión de los TBV en los EBV), sesgo (promedio de la diferencia entre EBV y TBV) y error cuadrático medio (MSE) para cada método de predicción. Los desvíos estándar (DS) se indican entre paréntesis.

Método de predicción	Exactitud	Inflación	Sesgo	MSE
PBLUP	0,51 (0,05)	0,98 (0,06)	–0,0003 (0,03)	0,206 (0,01)
ssGBLUP	0,72 (0,03)	0,89 (0,19)	0,2169 (0,04)	0,159 (0,03)
ssGBLUP_F	0,74 (0,02)	0,99 (0,04)	0,1167 (0,04)	0,141 (0,01)
ssGBLUP_M	0,74 (0,02)	0,94 (0,04)	0,0094 (0,03)	0,125 (0,01)

Los valores reportados en el cuadro corresponden a los promedios de las 20 réplicas de la simulación.

Los sesgos de predicción se presentan en el Cuadro 4.2 y en la Figura 4.1c. Tanto PBLUP como ssGBLUP_M no mostraron sesgo, mientras que ssGBLUP y ssGBLUP_F fueron sesgados hacia arriba. Dentro del segundo grupo, el sesgo fue mayor para ssGBLUP que para ssGBLUP_F, altamente influenciado por un *outlier.* De hecho, la mediana del sesgo fue muy similar para ambos. Para el caso de ssGBLUP_F el sesgo fue equivalente a aproximadamente 0,5 generaciones de mejoramiento genético o a 0,4 desvíos estándares genéticos. Finalmente, ssGBLUP_M presentó el menor MSE (aquel más cercano a cero) seguido por ssGBLUP_F, tal como se puede apreciar en el Cuadro 4.2.

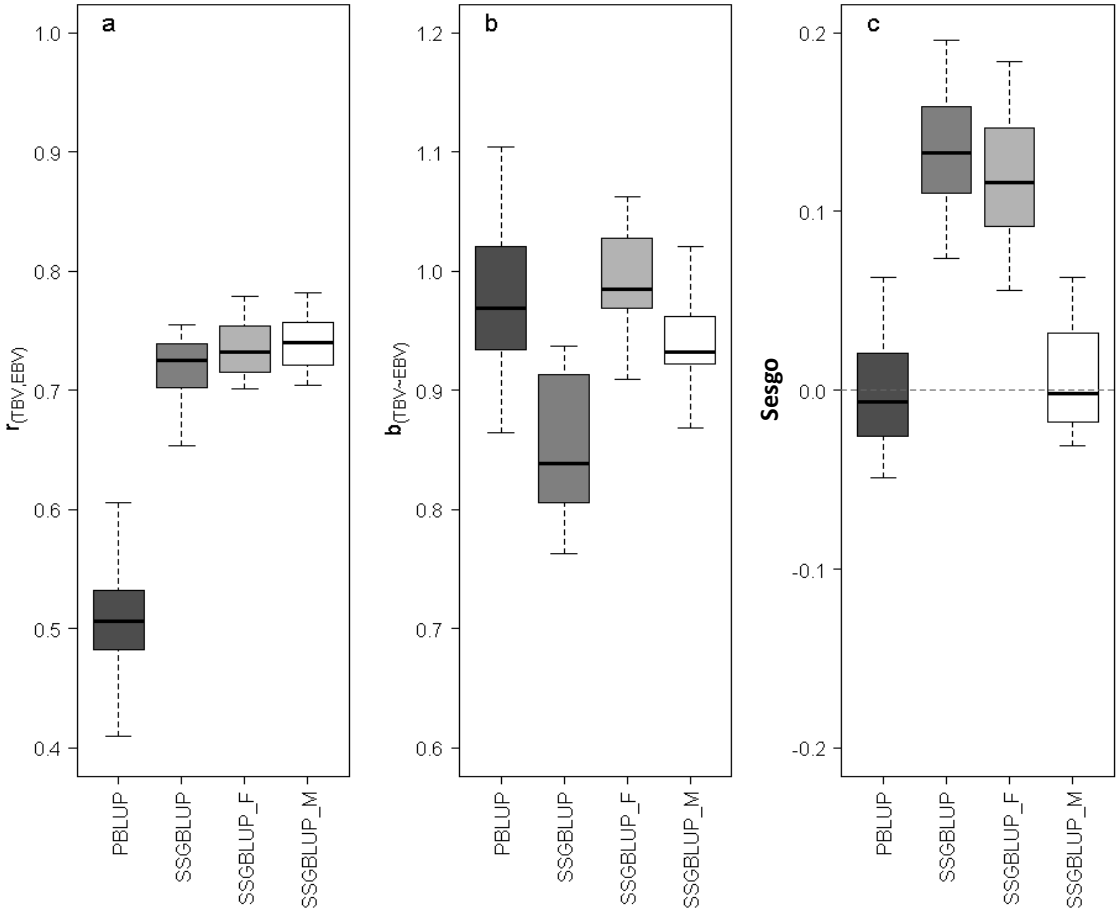


Figura 4.1. Diagramas de cajas y bigotes de a. correlación entre TBV y EBV para cada método (exactitud); b. coeficiente b_1 de la regresión de TBV en EBV (inflación); c. sesgo (promedio de la diferencia entre EBV y TBV).

Tal como se presenta en el Cuadro 4.3, emplear estimaciones de γ obtenidas por MM generó un impacto despreciable en términos predictivos, concretamente a nivel de la exactitud e inflación. Se tomaron sólo dos escenarios para evaluar el comportamiento del modelo ante cambios de γ, con el objetivo de cubrir dos situaciones extremas. Por un lado, se empleó la estimación de γ obtenida por MM, que representa el peor escenario de estimación según los resultados reportados en el capítulo 3. Por otra parte, se empleó el valor real del parámetro con el objetivo de trabajar en el escenario ideal y poder advertir cambios en términos predictivos. Tanto las exactitudes como los coeficientes b_1 no se vieron afectadas hasta el cuarto decimal. Los resultados sugieren que no es necesario estimar γ con precisión, dado que el impacto en términos predictivos es escasamente

importante. Cabe recordar aquí, tal como se mencionara en el capítulo 3, que dicho parámetro no solo es necesario para la predicción de los EBV incluyendo MF, sino que también permite caracterizar las poblaciones en términos de la variabilidad genética (F_{st}). Motivo por el cual es importante contar con una buena estimación del parámetro, más allá de los fines predictivos del mérito genético animal en la población bajo análisis.

Cuadro 4.3. Impacto del empleo de diferentes valores estimados de γ en términos predictivos a nivel de exactitud e inflación. Se empleó ssGBLUP_M con el parámetro γ estimado por el método de momentos (MM) y el verdadero valor conocido por simulación.

	Valor de γ	
	Estimado por MM	**Verdadero**
Exactitud	0,7412	0,7414
Inflación	0,9409	0,9400

4.3.2. Ranking de los candidatos a la selección según sus EBV predichos por cada metodología

Por otro lado, se compararon los métodos de predicción entre sí considerando las correlaciones de ranking entre EBV y TBV y entre los EBV predichos por las diferentes metodologías. Una correlación de ranking que toma valor uno implica que los mismos candidatos serán selectos independientemente de cuál de ambos métodos de predicción involucrados se emplee. Los resultados de dichas correlaciones se presentan en el Cuadro 4.4. Las correlaciones de ranking con los TBV fueron muy similares a las de Pearson presentadas en el Cuadro 4.2. En términos prácticos, las decisiones de selección variaron muy poco al emplear ssGBLUP, ssGBLUP_F o ssGBLUP_M. Nótese, que en el Cuadro 4.4 se presentan los valores de las correlaciones de ranking para los animales más jóvenes de la población (última generación) y no aborda las comparaciones entre generaciones (como podría ser animales más viejos vs. los más jóvenes), que son sensibles a los sesgos reportados en el Cuadro 4.2 (Winkelman *et al.*, 2015). Por ejemplo, en el caso de ssGBLUP_F, todos los animales jóvenes (candidatos a la selección) contarían con EBV sobreestimadas por 0,11 unidades. Esto los haría parecer mejores que aquellos padres probados que cuentan con alta exactitud (muy cercana a uno) y ausencia de sesgo. Según el esquema de mejoramiento genético animal, esta situación puede impulsar decisiones de selección subóptimas con impacto directo en la respuesta a la selección.

Cuadro 4.4 Correlaciones de Spearman entre TBV y EBV para cada método de predicción. Los desvíos estándar (DS) se indican entre paréntesis.

	EBV PBLUP	**EBV ssGBLUP**	**EBV ssGBLUP_F**	**EBV ssGBLUP_M**
TBV	0,49 (0,06)	0,71 (0,02)	0,72 (0,03)	0,73 (0,02)
EBV PBLUP		0,56 (0,05)	0,62 (0,04)	0,64 (0,04)
EBV ssGBLUP			0,99 (0,01)	0,98 (0,01)
EBV ssGBLUP_F				0,99 (0,002)

Los valores reportados en el cuadro corresponden a los promedios de 20 réplicas.

4.3.3. Estimación de los componentes de varianza

En la Figura 4.2 se observan las estimaciones de la heredabilidad obtenidas con tres de los cuatro métodos de predicción (PBLUP, ssGBLUP_F y ssGBLUP_M). No se reportan las estimaciones obtenidas empleando ssGBLUP dado que seis de las 20 réplicas no presentaron convergencia del procedimiento de estimación. Para alcanzar la convergencia en dichos casos fue necesario modificar la ponderación de la submatriz $\boldsymbol{A}_{22}^{-1}$ en $\boldsymbol{H}^{-1}$ tomando $\omega = 0,7$ en lugar de $\omega = 1$ (Tsuruta *et al.*, 2011). De todos modos, las estimaciones obtenidas fueron de mala calidad y este es el motivo por el cual no se reportan.

Tal como se observa en la Figura 4.2, las estimaciones fueron generalmente inferiores al valor real simulado para dicho parámetro (0,30), independientemente del método de predicción empleado. Las estimaciones más bajas se obtuvieron con ssGBLUP_F. Incluir un MF permitió mejorar la calidad de las mismas comparando contra ssGBLUP_F y redujo la variabilidad en relación a PBLUP. Además, al considerar *F* (ssGBLUP_F) o MF (ssGBLUP_M) fue posible llegar a convergencia en todas las réplicas.

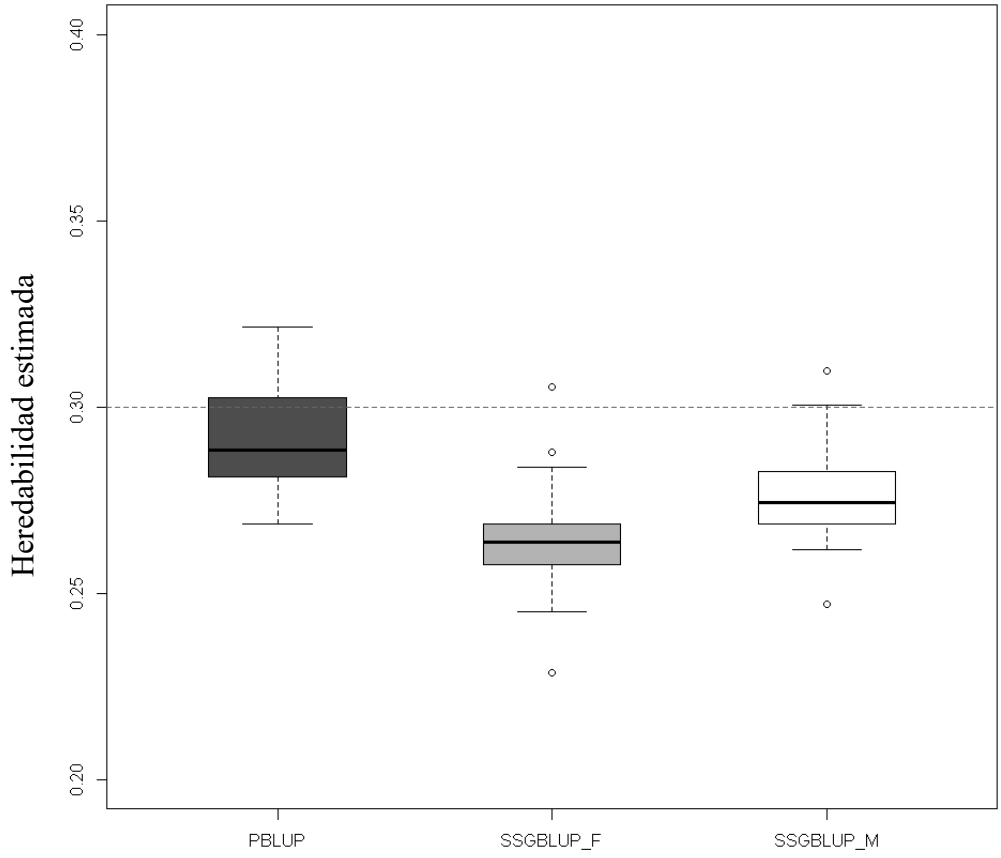


Figura 4.2. Diagramas de cajas y bigotes de la heredabilidad estimada empleando PBLUP, ssGBLUP_F y ssGBLUP_M considerando las 20 réplicas. La línea punteada indica el valor real simulado (h^2 = 0,30).

4.4. DISCUSION

Tanto en este capítulo como en los dos precedentes se abordó el problema de la falta de compatibilidad entre la matriz ***A*** y la ***G***. Powell *et al.* (2010) plantearon que tanto IBS como IBD son nociones compatibles debido a que ambas son medidas de identidad de genes. Ahora bien, al momento de combinar ambos tipos de relaciones, se presentan incompatibilidades (Vitezica *et al.*, 2011; Christensen *et al.*, 2012; Harris y Johnson, 2010; Meuwissen *et al.*, 2011). Legarra (2016) sugirió que, para comparar la varianza genética entre IBD, IBS u otras medidas de relación, es necesario tomar una referencia común. En el presente trabajo se empleó una referencia fija (***G*** construida como el producto cruzado de genotipos codificados como {−1,0,1}) y se adecuó ***A*** (pedigrí, IBD) para ajustarla a ***G*** (IBS, marcadores). Comparado con enfoques previos, emplear una referencia fija presenta la ventaja que las relaciones genómicas son inmutables (si se suman individuos genotipados

a la base de datos, las relaciones existentes no varían) y no dependen de la cantidad de información genealogía disponible, que por construcción es siempre limitada y, en mejoramiento genético animal, usualmente heterogénea. El enfoque de nuestro trabajo es similar a emplear IBS como una medida de identidad. Utilizamos una matriz de relaciones $\boldsymbol{G} = (\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})' / s = 2(\boldsymbol{M} - \boldsymbol{J})(\boldsymbol{M} - \boldsymbol{J})' / n$ dado que $s = n/2$, mientras que la matriz IBS es $\boldsymbol{G}_{\text{IBS}} = \boldsymbol{G}/2 + \boldsymbol{11}'$ (ver demostración en el apéndice). En GBLUP con estimación de componentes de varianza asociada cuando todos los animales se encuentran genotipados, emplear un modelo con $\boldsymbol{G}_{\text{IBS}}$ o la matriz $\boldsymbol{G}$ propuesta aquí producen EBV idénticos debido a que el término $1/2$ en $(\boldsymbol{G}/2)$ es absorbido en el componente de varianza, mientras que la constante $\boldsymbol{11}'$ es absorbida por la parte fija del modelo mixto (Legarra, 2016; Strandén y Christensen, 2011). De todos modos, es la matriz $\boldsymbol{G}$ presentada más arriba la que debe ser usada en ssGBLUP_M debido a que $\boldsymbol{G}_{\text{IBS}}$ no es compatible con las relaciones de pedigrí. En Fernando *et al.* (2014), el término de intercepción (efecto fijo) μ_g modela, análogamente a Vitezica *et al.* (2011), la diferencia entre los valores genéticos de los individuos en la base y aquellos genotipados. Consecuentemente, este término tiene un rol similar al de los MF.

4.4.1. Consecuencias de emplear metafundadores en las evaluaciones genómicas

Las estimaciones genómicas de los valores de cría son invariantes a la codificación cuando todos los individuos se encuentran genotipados (Strandén y Christensen, 2011). Ahora bien, este no es el caso cuando se combina la información genealógica y molecular, tal como ocurre en ssGBLUP. De este trabajo se desprende que, incluso contando con información completa de genealogía y una única población base, el empleo de los MF en ssGBLUP_M genera estimaciones de EBV con una inflación ligeramente mayor, pero con menos sesgo que aquellas obtenidas por ssGBLUP_F. En relación con la heredabilidad, con ssGBLUP_M se obtienen estimaciones casi insesgadas. El sesgo, definido como $\text{E}(\text{EBV} - \text{TBV})$ es comúnmente pasado por alto en predicciones genómicas. En un ejemplo de una evaluación genética sesgada, Henderson (1973) mostró que los padres de generaciones recientes se encontraban sub-evaluados con respecto a aquellos más viejos. La sobre-dispersión, a la que en literatura más reciente suele referirse en términos de sesgo (por ejemplo Mäntysaari *et al.*, 2010) también puede generar un gran impacto en la práctica (Sargolzaei *et al.*, 2012; Spelman *et al.*, 2010; Winkelman *et al.*, 2015) y la relación de compromiso entre sesgo y varianza aún requiere mayor estudio. Por ejemplo, Vitezica *et al.* (2011) encontraron que ssGBLUP_F es insesgado pero posee sobre-dispersión, lo cual probablemente depende de la estructura de los datos, incluyendo qué subconjunto de animales se encuentra genotipado.

Por su parte, el empleo de MF permite una definición clara de las relaciones genómicas ya que no dependen de la cantidad de información con la que se cuente en la genealogía ni de los cambios en las frecuencias alélicas al incorporar más información. Además, el parámetro de gran dimensión correspondiente a las frecuencias alélicas de la base es reemplazado por otro de menor dimensión (matriz **Γ**).

El pobre desempeño de ssGBLUP con respecto a ssGBLUP_F se da, probablemente, debido a la presencia de animales con altos niveles de *F* en la población simulada. Recuérdese que ssGBLUP ignora *F* a la hora de construir $\boldsymbol{A}^{-1}$. Esto se relaciona a la interpretación del parámetro ω, tal como fue empleado en otros trabajos en los que se utilizó ssGBLUP (Tsuruta *et al.*, 2011). Una aplicación de ssGBLUP para caracteres de tipo en Holstein presentó problemas de convergencia que se solucionaron al multiplicar $\boldsymbol{A}_{22}^{-1}$ por $\omega = 0,7$ al calcular la matriz $\boldsymbol{H}$ con una ganancia en exactitud de las predicciones (Tsuruta *et al.*, 2011). De todos modos, la naturaleza del parámetro ω es desconocida (Misztal *et al.*, 2013). En dicho trabajo, la inversa de la matriz de relaciones aditivas $\boldsymbol{A}^{-1}$ se construyó empleando las reglas de Henderson ignorando *F* (Mehrabani-Yeganeh *et al.*, 2000), mientras que en la submatriz $\boldsymbol{A}_{22}^{-1}$ se la tomó en consideración. Como consecuencia, los elementos de esta última matriz fueron muy grandes. Además, los animales genotipados no se encontraban relacionados en promedio en la matriz $\boldsymbol{G}$, pero si lo estaban en $\boldsymbol{A}_{22}$, aspecto que puede corregirse al escalar $\boldsymbol{G}$ tal como se presenta en Vitezica *et al.* (2011). Esto generó que los elementos de $\boldsymbol{A}_{22}^{-1}$ para los animales más jóvenes fuesen de mayor valor absoluto que los de $\boldsymbol{G}$. Ambos problemas pueden evitarse parcialmente al emplear un ponderador $\omega < 1$ para $\boldsymbol{A}_{22}^{-1}$. Cuando se construyó $\boldsymbol{A}^{-1}$ teniendo en cuenta *F*, el valor óptimo del ponderador en el análisis llevado a cabo en Holstein aumentó de 0,7 a 0,9 (Masuda, comunicación personal 2016). Ahora bien, el enfoque considerando a los MF permite alcanzar una solución más adecuada al problema. Además, basado en estas experiencias, $\boldsymbol{A}^{-1}$ debería construirse siempre considerando *F* para evitar casos patológicos que suelen ser infrecuentes pero muy problemáticos.

Capítulo 5

Estimación de la varianza aditiva y de dominancia empleando información genómica para caracteres de crecimiento en una población de bovinos de carne

Estimación de la varianza aditiva y de dominancia empleando información genómica para caracteres de crecimiento en una población de bovinos de carne

5.1. INTRODUCCIÓN

En las evaluaciones genéticas tradicionales la tendencia histórica fue a ignorar los efectos no aditivos (dominancia y epistasis) en el modelo. La misma persiste en la actualidad con los modelos genómicos, ampliamente utilizados en diversas especies de animales domésticos. En la mayoría de los casos sólo se consideran efectos aditivos a la hora de predecir el mérito genético de los candidatos a la selección. Los motivos que llevaron a ignorar los efectos no aditivos en los modelos son, principalmente, de índole operativo y práctico tal como los presentan Varona *et al.* (2018). Entre los primeros se destaca la falta de pedigríes informativos (con información de varias generaciones hacia atrás para cada individuo) y, en especial, de grandes familias de hermanos enteros. Contar con relaciones de "dominancia" tales como las de hermanos enteros (que involucran la probabilidad de genotipos idénticos de dos individuos en un locus) es de vital importancia (en especial para casos donde no se cuenta con información genómica) ya que constituye una de las relaciones más informativas para estos casos. Además, los cálculos al incorporar efectos no aditivos en el modelo suelen ser complejos y demandantes computacionalmente. Por otra parte, el hecho de que la varianza genética aditiva (*estadística*) incluya parte de los efectos no aditivos biológicos de los genes (Hill, 2010), contribuye a la idea de que no es necesario considerar directamente este tipo de efectos dentro del modelo, pese a que la varianza de dominancia (de la desviación de dominancia) puede explicar una parte significativa de la varianza genética total (Varona *et al.*, 2018). En consecuencia, en la actualidad existen pocas estimaciones de la varianza de los efectos no aditivos en poblaciones de animales domésticos. Entre las que se encuentran aquellas reportadas por Misztal *et al.* (1998), Fernández *et al.* (2017) y las de Nguyen y Nagyné-Kiszlinger (2016) para varias especies.

Por otro lado, incorporar los efectos no aditivos en el modelo puede tener ciertas ventajas (Varona *et al.*, 2018). Primero, puede aumentar la exactitud de predicción de los valores de cría y la respuesta a la selección (Toro y Varona, 2010; Aliloo *et al.*, 2016; Duenk *et al.*, 2017). Segundo, es posible definir apareamientos que permitan maximizar la performance productiva de los animales de la próxima generación considerando tanto el valor de cría como el valor genético total (valor aditivo más el valor de la desviación de dominancia) (Maki-Tanila, 2007; Toro y Varona, 2010; Aliloo *et al.*, 2017). Tercero, es posible sacar provecho de la variación genética no aditiva a través de la definición de cruzamientos o esquemas de apareamientos de animales puros (Maki-Tanila, 2007; Zeng *et al.*, 2013).

La oportunidad de sacar provecho de las ventajas mencionadas y la creciente disponibilidad de grandes volúmenes de información genómica renovaron el interés en la posibilidad de considerar los efectos no aditivos en los modelos empleados en las evaluaciones genómicas. Toro y Varona (2010), Su *et al.* (2012) y Vitezica *et al.* (2013) abordaron dicho desafío y propusieron diferentes alternativas para considerar los efectos no aditivos en los modelos valiéndose de la genómica como fuente de información adicional. Existen dos modos de parametrizar el modelo según cómo se definan los efectos genéticos del mismo. Por un lado, bajo el enfoque clásico de la genética cuantitativa, los valores de cría (efectos aditivos) están dados por los efectos de sustitución que incluyen tanto la parte aditiva como así también los efectos de dominancia biológicos de los genes. Las desviaciones de dominancia, por su parte, incluyen sólo una parte de los efectos biológicos de dominancia de los genes. Por otro lado, existe una parametrización alternativa que surge de considerar los valores genotípicos de los individuos. En este caso se definen los efectos aditivos y de dominancia como aquellos atribuibles a los efectos biológicos aditivos y de dominancia de los genes, respectivamente. Ambos enfoques serán abordados nuevamente más adelante a los efectos de detallar en mayor profundidad las características de cada uno y las diferencias entre ellos, dada su importancia a la hora de definir la parametrización a emplear y sus implicancias en función de las propiedades de cada uno.

Dentro de las diferentes metodologías que permiten considerar la información molecular en la predicción de mérito genético, se encuentra la posibilidad de construir una matriz de relaciones genómicas basada en la información de los marcadores y utilizarla como la matriz de estructura de las (co)varianzas de los valores de cría. Esta matriz se la conoce como $\boldsymbol{G}$ y existen diferentes metodologías para construirla condicionalmente a la información molecular. Ahora bien, nótese que también es posible valerse de la información de los marcadores moleculares para construir una matriz de relaciones de dominancia, conocida como $\boldsymbol{D}$. En este capítulo emplearemos aquella propuesta por Vitezica *et al.* (2013), que será descripta en detalle más adelante.

Los efectos no aditivos han sido incorporados en los modelos de evaluación genómica por varios autores en diferentes poblaciones animales. Entre estas se encuentran los bovinos lecheros (Sun *et al.*, 2013, Ertl *et al.*, 2014, Aliloo *et al.*, 2016, Jiang *et al.*, 2017), los cerdos (Esfandyari *et al.*, 2016; Xiang *et al.*, 2016), las ovejas (Moghaddar y van der Werf, 2017) y las gallinas ponedoras (Heidaritabar *et al.*, 2016). Los resultados reportados en estos casos fueron variables y, en algunos casos, ambiguos. Para el caso puntual de los bovinos de carne, se ha generado poca investigación reciente referida al tema. Esto puede atribuirse principalmente a la falta de grandes bases de datos con un gran número de animales con información genómica y fenotípica, tal como discute Varona *et al.* (2018). Recientemente, Bolormaa *et al.* (2015) evaluaron la factibilidad de incluir efectos no aditivos en el modelo empleando información genómica y fenotípica de animales de diferentes razas bovinas carniceras y sus cruzas. A tal fin consideraron varios caracteres pero sin enfocarse específicamente en aquellos de alto impacto en términos productivos y económicos para este tipo de sistemas de producción como son los de crecimiento y calidad

de res. Concretamente, evaluaron dos caracteres de crecimiento en el marco de un análisis general en el que se incluyeron varios caracteres de otros tipos. Por su parte, estudios previos en los que se emplea información de genealogía únicamente, sin recurrir a la genómica, como por ejemplo Rodriguez-Almeida *et al.* (1995), Gengler (1997), Misztal y Varona (1999) reportaron resultados ambiguos para varios caracteres de crecimiento en ganado de carne. Dichos resultados no permiten conocer con claridad que proporción de la varianza genética total es explicada por efectos no aditivos tales como las desviaciones de dominancia y/o las epistáticas.

En consecuencia, en este capítulo se aborda dicha problemática empleando una amplia base de datos de una población real de ganado de carne con información genómica y fenotípica. El análisis se centra en caracteres de crecimiento, dada su importancia en los sistemas productivos de ganado de carne. Se evalúa la conveniencia de incluir la dominancia a la hora de llevar a cabo las evaluaciones genómicas. A tal fin se compararán dos modelos: 1) sólo con efectos genéticos aditivos y 2) con efectos aditivos y de desviaciones de dominancia, siguiendo la metodología propuesta por Vitezica *et al.* (2013). El objetivo es obtener estimaciones de la varianza aditiva y de dominancia (varianza de las desviaciones de dominancia), considerando también la depresión consanguínea en el modelo, tal como será detallado más adelante.

El capítulo está organizado del siguiente modo. En primer lugar, se introducen brevemente conceptos centrales de la genética cuantitativa necesarios para comprender las diferencias entre las parametrizaciones de los modelos. Luego, se detallan las principales características de las mismas y sus propiedades con el objetivo de determinar cuál de las dos resulta más conveniente para el análisis propuesto en este capítulo. Finalmente, se describe el análisis llevado a cabo empleando datos de una población de bovinos de carne y se discuten los resultados obtenidos en función del modelo estadístico empleado.

5.2. MÉTODOS

5.2.1. Marco teórico

En esta sección se resumen ciertos conceptos básicos de la teoría de la genética cuantitativa que permiten comprender los modelos que se emplean más adelante y los resultados obtenidos.

5.2.1.1 Efectos aditivos y de desviaciones de dominancia

Bajo la teoría de la genética cuantitativa clásica, el valor de cría de un individuo involucra los efectos de sustitución de los genes (α), definidos del siguiente modo:

$$\alpha = a + d(q - p). \qquad [5.1]$$

Estos incluyen el efecto *biológico* aditivo (*a*), el efecto *biológico* de dominancia (*d*) de los genes y las frecuencias alélicas (*p* y *q*). Nótese que si no hubiera efectos de dominancia,

$$d = 0 \qquad [5.2]$$

y, al sustituir dicho valor en la expresión [5.1] se obtiene

$$\alpha = a\,. \qquad [5.3]$$

Es decir que en ausencia de dominancia, los efectos de sustitución se corresponden con los efectos aditivos biológicos, tal como se presenta en [5.3].

Ahora bien, si se considera un único locus con dos alelos (A_1 y A_2), puede definirse un efecto *biológico* para cada uno de los tres genotipos posibles, del siguiente modo:

$$\begin{aligned} A_1A_1 &= a \\ A_1A_2 &= d \\ A_2A_2 &= -a \end{aligned} \qquad [5.4]$$

Estos efectos se definen como desviaciones del punto medio entre los dos homocigotas (Falconer y Mackay, 1996). Por consiguiente, puede plantearse un modelo que permite considerar dichos efectos (aditivos y de dominancia) *biológicos* directamente. Dicha parametrización constituye en apariencia el modo más sencillo e intuitivo de considerar estos efectos y será descripto detalladamente más adelante, en la sección 5.2.1.4. En cambio, en el marco de la genética cuantitativa tradicional un objetivo primordial es calcular valores de cría, para lo cual se consideran efectos *estadísticos* en el modelo. En este enfoque, los valores de cría están dados por los efectos de sustitución de los genes que incluyen tanto la parte aditiva *biológica* como así también los efectos de dominancia *biológicos* (expresión [5.1]). Las desviaciones de dominancia, por otro lado, incluyen sólo una parte de los efectos *biológicos* de dominancia de los genes. Nótese que en este caso, contrario a lo que ocurre con el enfoque *biológico*, los valores de cría, las desviaciones de dominancia y sus componentes de varianza son resultados estadísticos definidos en un contexto poblacional (Hill *et al.*, 2008) y tienen uso directo en esquemas de selección.

Por tal motivo, antes de proceder a presentar los modelos, es necesario definir inequívocamente ambos enfoques para evitar confusiones o ambigüedades. A tal fin, adoptaremos los términos empleados por Vitezica *et al.* (2013) y Varona *et al.* (2018). Por un lado denotaremos como efectos de dominancia *biológicos* o *genotípicos* a las observaciones de la dominancia en términos de la acción génica a nivel de un locus individual (Hill *et al.*, 2008). Dicho de otro modo, en términos *biológicos*, la dominancia implica que el valor del heterocigota se encuentre desviado (por encima o por debajo) del punto medio de los genotipos homocigotas producto de la interacción entre genes dentro del mismo locus. Por otro lado, entenderemos por efectos *estadísticos* a los valores de cría (definidos por los efectos de sustitución) y a las desviaciones de dominancia (Hill *et al.*, 2008). De hecho, aun en casos donde la acción génica biológica es predominantemente dominante, la mayor parte de la variación genética se transforma en aditiva (Hill, 2017).

Los modelos a emplear en las evaluaciones genómicas considerando efectos aditivos y de dominancia pueden parametrizarse de dos modos distintos según dichos efectos sean considerados de modo *estadístico* (parametrización clásica) o *biológico* (parametrización alternativa). En las dos secciones que siguen se presentan las principales características de cada uno de ellos.

5.2.1.2 Modelo con efectos aditivos y de desviaciones de dominancia: parametrización clásica

De manera general hablaremos de efectos de marcadores, aunque en realidad los efectos son propios de los genes y los marcadores capturan esos efectos debido al desequilibrio de ligamiento entre ambos. El atribuir efectos a los marcadores es una hipótesis de trabajo común en desarrollos teóricos en un contexto genómico, y debe entenderse como una aproximación.

Existe la posibilidad de emplear un modelo genómico que incorpora a la dominancia, comparable al genético clásico (por ejemplo BLUP basado en genealogía). El mismo incluye valores de cría ($\boldsymbol{u}$) y desviaciones de dominancia ($\boldsymbol{v}$) del siguiente modo

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{u} + \boldsymbol{v} + \boldsymbol{e} \qquad [5.5]$$

donde

$$\boldsymbol{u} = \boldsymbol{Z}\boldsymbol{a}. \qquad [5.6]$$

La matriz $\boldsymbol{Z}$ se codifica tal como lo propone VanRaden (2008). Los elementos de dicha matriz pueden tomar diferentes valores según el genotipo del individuo *i* en el marcador *j*:

$$Z_{ij} = \begin{cases} (2-2p_j) \\ (1-2p_j) \text{ para los genotipos} \\ -2p_j \end{cases} \begin{cases} \mathrm{A_1A_1} \\ \mathrm{A_1A_2} \\ \mathrm{A_2A_2} \end{cases} \qquad [5.7]$$

Esto se debe a que, tal como lo presentan Falconer y Mackay (1996), los valores de cría de un individuo están dados del siguiente modo según el genotipo y las frecuencias alélicas (*p* y *q*):

$$u_{\mathrm{A_1A_1}} = 2q\alpha = (2-2p)\alpha$$

$$u_{\mathrm{A_1A_2}} = (q-p)\alpha = (1-2p)\alpha \qquad [5.8]$$

$$u_{\mathrm{A_2A_2}} = (-2p)\alpha$$

Por otro lado, el tercer término de la expresión [5.5] es

$$\boldsymbol{v} = \boldsymbol{W}\boldsymbol{d} \qquad [5.9]$$

donde los elementos de $\boldsymbol{W}$ para el individuo *i* en el marcador *j* pueden tomar algunos de los siguientes valores dependiendo del genotipo, tal como se detalla a continuación:

$$W_{ij} = \begin{cases} -2q_j^2 \\ 2p_jq_j \text{ para los genotipos} \\ -2p_j^2 \end{cases} \begin{cases} \mathrm{A_1A_1} \\ \mathrm{A_1A_2} \\ \mathrm{A_2A_2} \end{cases} \qquad [5.10]$$

Esto se debe a que la desviación de dominancia de un individuo (Falconer y Mackay, 1996) está dada del siguiente modo según el genotipo y las frecuencias alélicas:

$$v_{\mathrm{A_1A_1}} = -2q^2d$$

$$v_{\mathrm{A_1A_2}} = 2pqd \qquad [5.11]$$

$$v_{\mathrm{A_2A_2}} = -2p^2d\,.$$

Es importante recordar que en [5.6], el valor de cría (*u*) involucra a los efectos aditivos y de dominancia *biológicos* de los marcadores (*a* y *d*). Por su lado, la desviación de dominancia (*v*) incluye únicamente una porción de los efectos de dominancia biológicos de los marcadores (*d*).

Siguiendo a Falconer y Mackay (1996), las esperanzas de los valores de cría y de las desviaciones de dominancia son

$$E(u)=0, \quad E(v)=0. \tag{5.12}$$

Por un lado, la varianza genética aditiva es

$$\sigma_u^2 = 2pq\left[a+d(q-p)\right]^2 = 2pq\alpha^2. \tag{5.13}$$

Nótese, que en esta parametrización, la varianza aditiva involucra a la variación debida tanto a los efectos *biológicos* aditivos como a los de dominancia de los marcadores. En el caso de la varianza de dominancia está dada por la siguiente expresión:

$$\sigma_v^2 = \mathrm{E}(v^2) - [\mathrm{E}(v)]^2 = \mathrm{E}(v^2), \tag{5.14}$$

Al desarrollar la esperanza $E(v)$ en la expresión [5.12],

$$E(v) = p^2\left(-2q^2d\right) + 2pq\left(2pqd\right) + q^2\left(-2p^2d\right) = 0 \tag{5.15}$$

de modo que

$$\sigma_v^2 = p^2\left(-2q^2d\right)^2 + 2pq\left(2pqd\right)^2 + q^2\left(-2p^2d\right)^2 = 4p^2q^2d^2(q^2+2pq+p^2)$$

$$\sigma_v^2 = \left[2pqd\right]^2 \tag{5.16}$$

Nótese que la varianza de dominancia incluye una parte de la variación de los efectos *biológicos* de dominancia de los marcadores.

Al extenderlo a varios marcadores, y considerando como aleatorios a los efectos de los marcadores, se obtiene

$$\sigma_u^2 = \sum_{j=1}^{nsnp}(2p_jq_j)\sigma_{a0}^2 + \sum_{j=1}^{nsnp}(2p_jq_j(q_j-p_j)^2)\sigma_{d0}^2 \tag{5.17}$$

$$\sigma_v^2 = \sum_{j=1}^{nsnp}\left(2p_jq_j\right)^2\sigma_{d0}^2 \tag{5.18}$$

donde σ_{a0}^2 y σ_{d0}^2 son las varianzas de los marcadores para los componentes aditivos y dominantes, respectivamente.

La varianza genética total está dada por la siguiente expresión

$$\sigma_g^2 = \sigma_u^2 + \sigma_v^2, \tag{5.19}$$

donde el primer término corresponde a la varianza genética aditiva y el segundo, a la varianza de las desviaciones de dominancia. Nótese que la partición *estadística* de la varianza en componentes *estadísticos* debidos a la aditividad y dominancia no se corresponde con los efectos *biológicos* de los genes (Huang y Mackay, 2016) pero es útil a la hora de predecir y llevar a cabo decisiones de selección (Vitezica *et al.*, 2013). Aun cuando los genes poseen una acción biológica dominante, esta variación es capturada principalmente por la varianza genética aditiva (Hill, 2017).

La parametrización *estadística* o clásica asume EHW y LE. Al asumir que los efectos de los marcadores no están correlacionados entre sí y son aleatorios (*a* y *d*), puede extenderse a varios loci (VanRaden, 2008, Gianola *et al.*, 2009) y se obtiene

$$\text{Var}(\boldsymbol{u}) = \frac{\boldsymbol{ZZ'}}{2\sum_{j=1}^{nsnp} p_j q_j}\sigma_u^2 = \boldsymbol{G}\sigma_u^2 \qquad [5.20]$$

Dicha matriz se corresponde con la ***G*** clásica de relaciones aditivas genómicas de GBLUP según VanRaden (2008). La varianza aditiva (σ_u^2) está dada por la expresión [5.17]. Por otro lado, para el caso de las desviaciones de dominancia (***v***), las (co)varianzas están dadas por

$$\text{Var}(\boldsymbol{v}) = \boldsymbol{WW'}\sigma_{d0}^2 \qquad [5.21]$$

Al reemplazar σ_{d0}^2 en [5.21] por $\sigma_{d0}^2 = \sigma_v^2 \Big/ \sum_{j=1}^{nsnp}\left(2p_j q_j\right)^2$ (que surge de despejar σ_{d0}^2 de la expresión [5.18]) se obtiene la matriz ***D*** de relaciones genómicas de desviaciones de dominancia, llegando a la siguiente expresión tal como se presenta en Vitezica *et al.* (2013).

$$\text{Var}(\boldsymbol{v}) = \frac{\boldsymbol{WW'}}{\sum_{j=1}^{nsnp}\left(2p_j q_j\right)^2}\sigma_v^2 = \boldsymbol{D}\sigma_v^2 \qquad [5.22]$$

La matriz ***D*** posee algunas características similares a la matriz ***G***. Por ejemplo, en la población base en EHW, el promedio de los elementos diagonales de ***G*** es uno, mientras que el de los de fuera de la diagonal es cero. Bajo estos mismos supuestos, lo mismo ocurre con los elementos de la matriz ***D.***

En el modelo clásico los efectos son ortogonales al asumir LE y EHW. Es decir que la definición (y por tanto, en general, la estimación) de un efecto no se ve afectada por la presencia de otros en el modelo. Esto permite contar con una partición ortogonal de las varianzas debido a que los efectos de sustitución contribuyen a la varianza aditiva y las desviaciones de dominancia contribuyen a la varianza de dominancia. La ortogonalidad implica que no existen covarianzas entre los dichos efectos y, como consecuencia, al incluir

nuevos efectos en el modelo las estimaciones no varían. De hecho, al pasar de un modelo que sólo incluye efectos aditivos a otro con efectos aditivos y de desviaciones de dominancia, las estimaciones de la varianza aditiva serán similares. Es decir que el modelo presentado permite una descomposición ortogonal de las varianzas genéticas en cualquier población bajo EHW y, en consecuencia, obtener estimaciones de los componentes de varianza comparables a los obtenidos al trabajar con información genealógica (Vitezica *et al.*, 2013).

5.2.1.3. Modelo Animal GDBLUP

Una vez definida la matriz de relaciones de dominancia genómicas presentada en la sección anterior (expresión [5.22]) es posible emplearla directamente en las predicciones genómicas GBLUP en el marco de las MME. A tal fin puede plantearse el siguiente modelo:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{Zv} + \boldsymbol{e} \qquad [5.23]$$

donde $\boldsymbol{Z}$ es una matriz de incidencia que relaciona el fenotipo con los valores de cría y las desviaciones de dominancia. Tomando

$$\text{Var}(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2, \ \text{Var}(\boldsymbol{v}) = \boldsymbol{D}\sigma_v^2 \ \text{ y } \ \text{Var}(\boldsymbol{e}) = \boldsymbol{R} \qquad [5.24]$$

y asumiendo normalidad multivariada, las MME toman la siguiente forma

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\sigma_u^{-2} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1}\sigma_v^{-2} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix} \qquad [5.25]$$

Estas ecuaciones son idénticas a las del modelo animal clásico, excepto por el hecho de emplear relaciones genómicas en las matrices $\boldsymbol{G}$ y $\boldsymbol{D}$, en lugar de las genealógicas (de la matriz $\boldsymbol{A}$). Los valores de cría y las desviaciones de dominancia pueden predecirse empleando estas MME.

5.2.1.4. Modelo con efectos aditivos y de dominancia: parametrización alternativa. Modelo genotípico.

Existe una parametrización alternativa a la presentada en la sección precedente y corresponde al modelo mencionado en la sección 5.2.1.1. Tal como se mencionara

anteriormente, el mismo incluye los efectos *biológicos* aditivos y de dominancia del gen (o marcador) y toma la siguiente forma

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Ta} + \boldsymbol{Xd} + \boldsymbol{e} \qquad [5.26]$$

Donde el vector $\boldsymbol{y}$ contiene los fenotipos, μ corresponde a la media de la población, $\mathbf{1}$ es un vector de unos, $\boldsymbol{a}$ es un vector de efectos aditivos de los *n* marcadores, $\boldsymbol{d}$ es un vector de efectos de dominancia para cada uno de los *n* marcadores considerados en el análisis y $\boldsymbol{e}$, un vector de residuos. Los elementos de la matriz $\boldsymbol{T}$ pueden tomar los siguientes valores según los genotipos de cada marcador

$$T_{ij} = \begin{cases} 1 \\ 0 \text{ para los genotipos} \\ -1 \end{cases} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases} \qquad [5.27]$$

Por su parte, los elementos de la matriz $\boldsymbol{X}$ pueden tomar los siguientes valores según los genotipos de los marcadores

$$X_{ij} = \begin{cases} 0 \\ 1 \text{ para los genotipos} \\ 0 \end{cases} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases} \qquad [5.28]$$

A diferencia del modelo presentado en [5.5], donde los efectos considerados son *estadísticos*, en [5.26] el significado es *biológico* para cada uno de sus *n* marcadores tal como lo proponen Toro y Varona (2010).

Este modelo se basa en genotipos observados, particularmente en heterocigotas, de modo que puede llamarse modelo *genotípico*. La extensión a un modelo animal fue propuesta por Su *et al.* (2012) y define

$$\boldsymbol{u}^* = \boldsymbol{Ta} \text{ y } \boldsymbol{v}^* = \boldsymbol{Xd}\,. \qquad [5.29]$$

Donde $\boldsymbol{u}^*$ y $\boldsymbol{v}^*$ son los efectos (propios de los individuos) aditivos y de dominancia *genotípicos*. Es importante destacar que los elementos de $\boldsymbol{u}^*$ no son valores de cría dado que no están definidos en términos de efectos de sustitución, sino que corresponden a la parte atribuible a los efectos aditivos *biológicos* de los marcadores. La matriz de incidencia $\boldsymbol{T}$ corresponde a la matriz $\boldsymbol{Z}$ empleada en el modelo clásico (expresión [5.5]). Ahora bien, la matriz $\boldsymbol{X}$ (en [5.26]) y la $\boldsymbol{W}$ (en [5.9]) son diferentes. La última es empleada en el modelo clásico para las desviaciones de dominancia.

En este caso, la varianza del valor aditivo genotípico puede obtenerse del siguiente modo

$$\sigma^2_{u^*} = \mathrm{E}(u^{*2}) - [\mathrm{E}(u^*)]^2. \quad [5.30]$$

La misma expresión puede emplearse para la varianza de los valores de dominancia genotípicos ($\sigma^2_{v^*}$). Entonces

$$\sigma^2_{u^*} = \sum 2 p_j q_j \sigma^2_a \quad [5.31]$$

y

$$\sigma^2_{v^*} = \sum 2 p_j q_j (1 - 2 p_j q_j) \sigma^2_d \quad [5.32]$$

Nótese que estas expresiones son diferentes a aquellas empleadas para el caso del modelo clásico de la sección 5.2.1.2 (expresiones [5.17] y [5.18]). Las varianzas $\sigma^2_{u^*}$ y $\sigma^2_{v^*}$ estimadas bajo un modelo *genotípico*, tal como propone Su *et al.* (2012), no son comparables directamente a aquellas obtenidas empleando el modelo clásico.

5.2.1.5. Diferencias entre ambas parametrizaciones e impactos en términos prácticos

Vitezica *et al.* (2013) mostraron que las matrices de relaciones de dominancia (***D***) son diferentes entre el modelo clásico (*estadístico*) y el modelo *genotípico*. El modelo clásico, en términos de valores de cría, efectos de sustitución, desviaciones de dominancia y varianzas genéticas, es más adecuado para llevar a cabo selección. Los únicos componentes de varianza comparables a aquellos basados en genealogía son σ^2_u and σ^2_v obtenidos al emplear el modelo genómico *estadístico* (Vitezica *et al.*, 2013). Las estimaciones del modelo *genotípico* no son varianzas genéticas y no pueden compararse con las varianzas calculadas utilizando la genealogía (Vitezica *et al.*, 2013).

Además, el modelo *estadístico* propuesto por Vitezica *et al.* (2013) es ortogonal (Vitezica *et al*., 2017) . Es decir que al introducir nuevos efectos genéticos en el modelo, no se esperan cambios en las estimaciones respecto de las previas (calculadas antes del incluir el nuevo efecto). Contrario a esto, el modelo genotípico de Su *et al.* (2012) no lo es, al incorporar la dominancia en el modelo pueden observarse cambios importantes en los valores aditivos obtenidos y en las estimaciones de las varianzas. En consecuencia, consideramos que el modelo clásico resulta más conveniente para llevar a cabo el análisis propuesto en este capítulo.

5.2.1.6. Depresión consanguínea

Tanto la depresión consanguínea como la heterosis pueden explicarse por dominancia direccional (Lynch y Walsh, 1998). En caracteres con depresión por consanguinidad o heterosis es de esperar observar un porcentaje mayor de efectos de dominancia *biológicos* positivos que negativos. En estos casos, la media de los efectos de dominancia de los marcadores (*d* en [5.11]) es distinta de cero, escenario que contrasta fuertemente con los supuestos de los modelos tradicionales: medias de ***a*** y ***d*** iguales a cero. Con el objetivo de considerar este último caso, Xiang *et al.* (2016) recomendaron incluir la consanguinidad genómica como una covariable en el modelo. Esto se debe a que mostraron que de este modo es posible tomar en cuenta la dominancia direccional de una manera que explicaría asimismo la depresión consanguínea. Por lo tanto, la consanguinidad genómica debe incluirse siempre en el modelo con el objetivo de obtener una estimación correcta de la varianza de dominancia (de Boer y Hoeschele, 1993; Aliloo *et al.*, 2016). Si la consanguinidad no es considerada dicha varianza es sobreestimada, tal como mostraron Xiang *et al.* (2016) y Aliloo *et al.* (2016) con datos reales. Además, la estimación de depresión por consanguinidad tiene interés *per se* en tanto que es un fenómeno biológico relevante para la cría y manejo del rodeo.

5.2.2. Análisis de datos reales

5.2.2.1 Descripción del archivo de datos

Los datos empleados en el análisis provienen de una población real de bovinos de carne. Para el análisis propuesto en este capítulo se emplearon los datos de tres caracteres de crecimiento: peso al nacer (PN), peso al destete (PD) y ganancia de peso post destete (GPD). Para el análisis se emplearon 19.375 animales, todos machos con información fenotípica para al menos uno de los tres caracteres. Los animales se encontraban genotipados con el panel BovineSNP50k v2 BeadChip (Illumina) de 54.609 marcadores. Se llevó a cabo un control de los registros fenotípicos con el objetivo de verificar que en todos los casos tomasen valores biológicamente posibles para cada carácter. Por otro lado, aquellos grupos de contemporáneos con menos de 10 individuos fueron eliminados del análisis y sólo se tomó en cuenta la información fenotípica de aquellos animales que contaban con información genómica. Como consecuencia, luego de la edición de los datos, se contaba con los genotipos de 19.357 animales para la construcción de las matrices de relaciones genómicas (***G*** y ***D***). Previo al cálculo de las mismas, se llevó a cabo un control de calidad de los datos genómicos empleando el software QCf90 (disponible en http://nce.ads.uga.edu/wiki/). Se filtró por *call rate* a nivel de marcador y animal, frecuencia del alelo en menor proporción (MAF, del inglés *minor allelle frequency*) con un valor umbral de 0,01 y EHW. Además, aquellos SNPs ubicados en los cromosomas sexuales fueron eliminados al igual que aquellos con posición desconocida. Como resultado

de esta edición se obtuvieron 39.245 SNPs restantes que fueron empleados para calcular las matrices ***G*** y ***D***. El Cuadro 5.1 presenta un breve resumen de las principales características de la base de datos empleada para el análisis.

Cuadro 5.1. Descripción de la base de datos fenotípicos en la que se cuenta con registros para tres caracteres de crecimiento: peso al nacer (PN), peso al destete (PD) y ganancia de peso post destete (GPD).

	PN	**PD**	**GPD**
Nº Animales con registros fenotípicos	19.375	19.345	14.767
Nº de grupos de contemporáneos	718	730	532
Promedio fenotípico (kg)	36,29	312,07	233,60
Desvío Estándar (kg)	3,63	40,37	48,08

5.2.2.2. Modelos estadísticos

Los fenotipos fueron analizados empleando un modelo GBLUP univariado. Los componentes de varianza se estimaron empleando dos modelos: 1) MG: sólo con efectos aditivos (expresión [5.33]) y 2) MGD: con efectos aditivos y de desviaciones de dominancia (expresión [5.34]). Luego de realizar análisis exploratorios de la base de datos se encontró que casi la totalidad de las madres (15,579) contaban con un solo hijo, generando que el promedio de hijos por vaca fuera de 1,10. En consecuencia los efectos maternos se encontraban completamente confundidos con el residuo en el modelo, motivo por el cual no fueron incluidos en el análisis de PN y PD.

Los modelos MG y MGD empleados en el análisis son

$$\text{MG: } \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{f}\,b + \boldsymbol{Zu} + \boldsymbol{e} \qquad [5.33]$$

$$\text{MGD: } \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{f}\,b + \boldsymbol{Zu} + \boldsymbol{Zv} + \boldsymbol{e} \qquad [5.34]$$

Donde ***y*** es un vector que contiene a los fenotipos observados para cada animal y para cada carácter, ***X*** es una matriz de diseño que permite relacionar a los fenotipos con los efectos fijos (grupos de contemporáneos), ***β*** es un vector de efectos fijos (grupos de contemporáneos), el vector ***f*** contiene los coeficientes de consanguinidad genómicos de los animales calculados como la proporción de loci en estado homocigota para cada individuo,

siguiendo a Silió *et al.* (2013) y Xiang *et al.* (2016). Por su parte, *b* es el parámetro de la depresión consanguínea, ***Z*** es una matriz de incidencia que relaciona el fenotipo con los valores de cría y las desviaciones de dominancia, ***u*** es un vector con los valores de cría, ***v*** es un vector de desviaciones de dominancia y ***e*** es un vector de residuos.

La matriz de relaciones aditivas genómica fue calculada siguiendo a VanRaden (2008), empleando la siguiente expresión:

$$\boldsymbol{G} = \frac{\boldsymbol{MM'}}{2\sum_{k=1}^{m} p_k q_k} \qquad [5.35]$$

Donde $\boldsymbol{M}$ es una matriz cuya dimension está dada por el número de animales (*n*) y por el número de SNPs (*m*), con elementos $(2-2p_k)$, $(1-2p_k)$ y $-2p_k$, según el genotipo A_1A_1, A_1A_2 y A_2A_2, respectivamente. En este caso, p_k corresponde a la frecuencia alélica del alelo A_1 para el marcador k y $q_k = 1 - p_k$.

Por su parte, la matriz de relaciones de dominancia ***D*** fue construida siguiendo a Vitezica *et al.* (2013), empleando la siguiente expresión:

$$\boldsymbol{D} = \frac{\boldsymbol{WW'}}{\sum_{k=1}^{m} (2p_k q_k)^2} \qquad [5.36]$$

Donde $\boldsymbol{W}$ posee la misma dimensión que $\boldsymbol{M}$, con elementos $-2q_k^2$, $2p_k q_k$ y $-2p_k^2$ según el genotipo del marcador: A_1A_1, A_1A_2 y A_2A_2, respectivamente.

5.2.2.3. Estimación de los componentes de varianza y comparación de modelos

La estimación de componentes de varianza se llevó a cabo empleando un muestreo de Gibbs y por máxima verosimilitud restringida (REML). Para cada carácter se generaron 200.000 muestreos usando software GIBBS2F90 y REMLF90 (disponibles en http://nce.ads.uga.edu/wiki/). Los parámetros iniciales de REML se obtuvieron a partir de las estimas de muestreo de Gibbs y solo 1-2 iteraciones fueron necesarias. En el muestreo de Gibbs, las primeras 10.000 muestras fueron descartadas dado que correspondían al período de calentamiento (*burn-in*). Se almacenó una de cada 10 muestras de las 190.000 restantes. La convergencia se verificó empleando un análisis visual de las cadenas y su variabilidad. Como criterio de bondad de ajuste se empleó, a partir de los resultados de REML, el test de cociente de verosimilitudes (del inglés *maximum likelihood ratio test*) y el criterio Akaike (AIC) para evaluar y comparar la bondad de ajuste de los modelos MG y MGD descriptos en detalle en la sección 5.2.2.2. Para la primera prueba, los valores χ^2 se calcularon empleando la siguiente expresión:

$$\chi^2 = -2logL_{MG} + 2logL_{MGD} \quad [5.37]$$

En la que el primer término involucra a la verosimilitud del modelo MG y el segundo, la del modelo MGD. Luego se obtuvieron los valores *p* de una distribución χ^2 mixta de cero y un grados de libertad tal como fue propuesto por Visscher (2006).

5.3. RESULTADOS

En el Cuadro 5.2 se presentan las estimaciones de los componentes de varianza obtenidos para cada uno de los tres caracteres de crecimiento empleando los modelos MG y MGD. Se obtuvieron resultados similares empleando el muestreo de Gibbs y REML, tal como era de esperar teóricamente. Nótese que para los tres caracteres la estimación de la varianza genética aditiva (σ_A^2) no se vio alterada al incorporar la dominancia en el modelo, tal como era de esperarse por emplear un modelo que permite particionar la varianza genética ortogonalmente, como lo presentaron Vitezica *et al.* (2017). Nótese que lo mismo ocurre con la heredabilidad, la misma no varía según el modelo empleado.

En relación a la varianza de las desviaciones de dominancia (σ_D^2), se desprende del Cuadro 5.2 que para todos los caracteres analizados la misma fue relativamente baja. En todos los casos no superó el 10% del valor de la varianza genética aditiva, lo que permite afirmar que hay evidencia de una reducida variación genética no aditiva debida a la dominancia para los caracteres de crecimiento en la población de bovinos de carne estudiada.

Cuadro 5.2. Estimaciones de los componentes de varianza y desvío estándar para cada uno de los caracteres de crecimiento empleando dos modelos: MG: únicamente con efectos aditivos y MGD: con efectos aditivos y de desviaciones de dominancia. Los valores de DS se presentan entre paréntesis.

Carácter	**PN**		**PD**		**GPD**	
Modelo	**MG**	**MGD**	**MG**	**MGD**	**MG**	**MGD**
σ_A^2	6,27 (0,33)	6,28 (0,33)	222,75 (14,61)	223,55 (14,82)	270,76 (20,42)	270,30 (21,94)
σ_D^2	-	0,18 (0,15)	-	10,02 (4,98)	-	21,68 (10,95)
h_A^2	0,25	0,25	0,16	0,16	0,16	0,16
h_D^2	-	0,01	-	0,01	-	0,01
σ_D^2/σ_A^2	-	0,03	-	0,04	-	0,08
σ_e^2	18,82 (0,24)	18,65 (0,28)	1186,28 (14,26)	1176,88 (14,86)	1388,81 (19,87)	1369,01 (26,00)

En el Cuadro 5.3 se presenta el promedio de la consanguinidad genómica calculada como la proporción de marcadores en estado homocigotas por individuo, siguiendo a Silió *et al.* (2013) y, en la Figura 5.1 se observa la distribución de dichos coeficientes. Por su parte, las estimaciones de la depresión consanguínea presentadas en el Cuadro 5.3 se encuentran expresadas en términos del cambio en la media fenotípica para cada carácter por cada 10% de incremento en la consanguinidad. Es así que para el caso del carácter PN se espera una reducción de 0,480 kg en la media del carácter por cada 10% de incremento en la consanguinidad. Siguiendo el mismo razonamiento, se espera una reducción de aproximadamente 10,225 kg para PD y 10,695 kg para GPD en la media de cada carácter por cada 10% de incremento en la consanguinidad.

Cuadro 5.3. Estimaciones de la consanguinidad genómica (*f*) y de la depresión consanguínea (*b*) para los tres caracteres de crecimiento analizados, empleando dos modelos (MG y MGD). Los valores de la depresión consanguínea están expresados en kg e indican el cambio en la media fenotípica para cada carácter por cada 10% de incremento en la consanguinidad. Los valores de DS se presentan entre paréntesis.

	PN	**PD**	**GPD**
Consanguinidad (*f*)		0,634 (0,013)	
MG (*b*)	–0,480 (0,195)	–10,225 (1,510)	–10,695 (1,900)
MGD (*b*)	–0,476 (0,207)	–9,998 (1,588)	–10,230 (2,036)

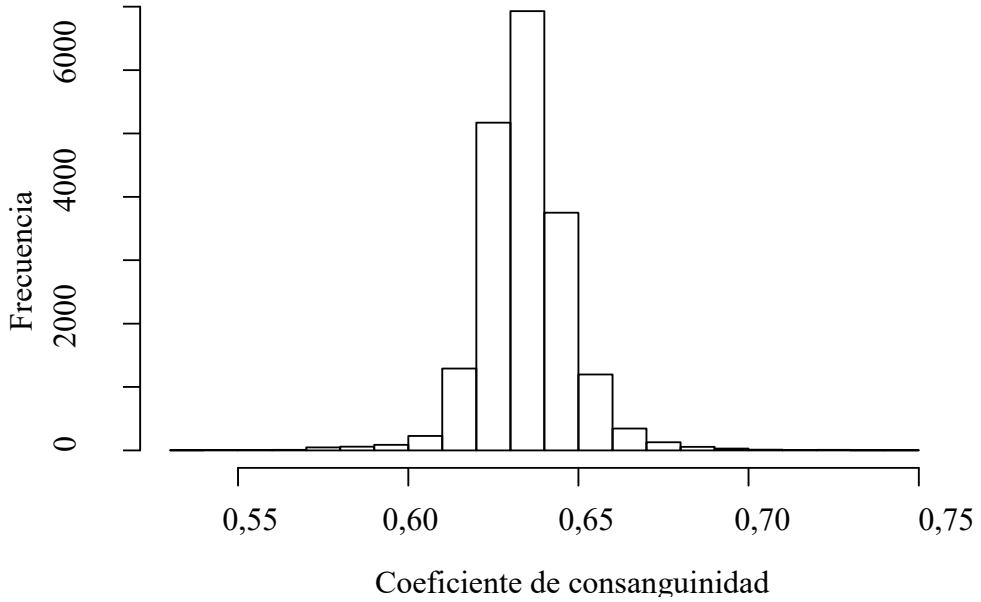


Figura 5.1. Coeficientes de consanguinidad genómica (*f*) calculados siguiendo a Silió *et al.* (2013).

Por otro lado, se calcularon los promedios de los elementos en la diagonal y fuera de la misma de las matrices de (co)varianzas ***G*** y ***D***, como instancia de control. Los valores se presentan en el Cuadro 5.4. Nótese que los mismos toman valores cercanos a los esperados para una población en EHW, tal como se presentó en la sección 5.2.1.2. Dada la similitud en los valores de DS para los valores fuera de la diagonal de ambas matrices, puede concluirse que no existe diferencia entre ellas en términos de informatividad.

Cuadro 5.4. Promedio y desvío estándar (DS) de los elementos en la diagonal y fuera de ella de las matrices *G* y *D*. Los DS se presentan entre paréntesis.

	G	***D***
Promedio de los elementos diagonales	1,02 (0,04)	1,01 (0,04)
Promedio de los elementos fuera de la diagonal	-5,24 e-05 (1,06)	0,02 (1,05)

Por otro lado, también se evaluó la bondad de ajuste de los modelos empleando el test del cociente de verosimilitudes, cuyos resultados se presentan en el Cuadro 5.5. Para los tres caracteres de crecimiento analizados, el modelo con mejor ajuste fue el MG. Dichos resultados muestran que el hecho de incluir la dominancia en el modelo para PN, PD y GPD no mejora el ajuste de los datos. Similares resultados se obtuvieron al evaluar los modelos empleando el Criterio de Información de Akaike (AIC), los valores se presentan en el Cuadro 5.6. Para todos los caracteres el modelo con menor (o igual) valor de AIC fue MG. Es importante mencionar que la distribución χ^2 toma únicamente valores no negativos debido a que está dada por la sumatoria de valores al cuadrado. Nótese que para PD y GPD el estadístico χ^2 toma valores inferiores a cero pero muy cercanos a dicho valor. Esto pudo deberse al empleo de aproximaciones (redondeo numérico).

Cuadro 5.5. Bondad de ajuste de los modelos MG y MGD y prueba de bondad de ajuste (valor χ^2 y *p* valor) entre los modelos MG y MGD para cada uno de los tres caracteres de crecimiento (PN, PD y GPD).

carácter	**-2 logL**			
	MG	**MGD**	χ^2	**P-valor**
PN	127.529,78	127.527,76	2,02	0,08
PD	203.515,41	203.515,89	-0,49	1,00
GPD	163.439,91	163.441,62	-1,71	1,00

Cuadro 5.6. **Valores de AIC para los modelos MG y MGD para cada uno de los tres caracteres de crecimiento (PN, PD y GPD).**

carácter	**AIC**	
	MG	**MGD**
PN	127.533,78	127.533,75
PD	203.519,41	203.521,89
GPD	163.443,91	163.447,62

5.4. DISCUSIÓN

En este capítulo se estimaron los componentes de varianza para tres caracteres de crecimiento empleando una extensa base de datos (genómicos y genotípicos) de una población bovina de carne. A tal fin se emplearon dos modelos bajo el enfoque clásico (MG y MGD), cuya principal diferencia estuvo dada por la incorporación de las desviaciones de dominancia en el último caso. Los objetivos principales fueron obtener estimaciones de la varianza aditiva y de dominancia para caracteres de importancia productiva en ganado de carne y comparar las estimaciones obtenidas empleando ambos modelos, así como también el ajuste de cada uno de ellos para determinar la conveniencia de incorporar la dominancia en los modelos de evaluación genómica para este tipo de caracteres. En relación al primer objetivo, los valores obtenidos para la varianza aditiva fueron consistentes con aquellos previamente reportados para la misma población. Ahora bien, en el caso de la varianza de dominancia, se obtuvieron valores bajos, inferiores al 10% de las varianzas aditivas. Los valores variaron dentro de un rango comprendido entre el 3 al 8 % de la varianza aditiva. Estos resultados permiten suponer que existe una reducida variación genética no aditiva debida a la dominancia para los caracteres de crecimiento en la población bajo estudio.

En relación al segundo objetivo mencionado, al pasar de un modelo que sólo incluye efectos aditivos (MG) a otro que incorpora la dominancia (MGD), las estimaciones de la varianza aditiva no se modificaron, tal como es esperado debido a la partición ortogonal de la varianza genética del modelo (Vitezica *et al.,* 2017). Por su parte al evaluar el ajuste de ambos modelos y compararlos entre sí, se observó que en ninguno de los casos se observaron mejorías en este aspecto al incorporar la dominancia.

Varona *et al.* (2018) discute acerca de los obstáculos que aún hoy existen para lograr implementar de modo estándar la incorporación de los efectos no aditivos en las evaluaciones genómicas. Entre estos obstáculos destaca la falta de una evaluación seria de los modelos que permiten incorporar dichos efectos, debido a la falta de bases de datos extensas que involucren un gran número de animales con información genómica y fenotípica. En este sentido, el análisis llevado a cabo en este capítulo constituye un aporte relevante ya que se empleó una base de datos de estas características en ganado de carne. Los resultados de este capítulo contribuyen a generar un antecedente de importancia en este aspecto y son complementarios a aquellos reportados por Bolormaa *et al.,* (2015) para otras poblaciones de bovinos de carne. Además permiten generar recomendaciones a la hora de querer incorporar la dominancia en la evaluación genómica de caracteres de crecimiento.

Por otro lado, si bien la base de datos empleada para el análisis contaba con un gran número de animales con registros fenotípicos e información genómica, la informatividad de la misma no fue óptima. De modo que es importante considerar este aspecto a la hora de analizar las estimaciones obtenidas. Por ejemplo, dada la baja proporción de hijos por madre, los efectos maternos se encontraban completamente confundidos, de modo que no pudieron considerarse dentro del modelo. Para ganar en informatividad sería necesario

contar con más animales con registros, buscando aumentar el número de hijos por madre. Esta estrategia podría contribuir, también a incrementar el número de relaciones de hermanos enteros, una de las más informativas en lo que a dominancia refiere.

Tal como fuera mencionado anteriormente, en la literatura existen pocas estimaciones de la varianza de dominancia reportadas en ganado de carne. Existen algunos trabajos (por ejemplo, Misztal y Varona, 1998; Gengler, 1997; Rodriguez-Almeida *et al.*, 1995) que reportan estimaciones basadas en información genealógica, mientras que aquellas incorporando información genómica son muy escasas. Bolormaa *et al.* (2015) obtuvieron estimaciones incorporando información genómica para un conjunto de bovinos de carne de diferentes razas. En dicho trabajo se analizaron diferentes caracteres, sin concentrarse específicamente en los de crecimiento de relevancia económica (los que más comúnmente se incorporan en las evaluaciones genéticas) como son PN, PD y GPD. Por su parte, los resultados de los trabajos basados en información genealógica de bovinos de carne son muy variables. Por ejemplo, para el carácter GPD los valores de h_d^2 reportados para razas carniceras como Limousin oscilan entre 0,103 y 0,184 (Misztal y Varona, 1999; Gengler 1997), valores muy superiores a los obtenidos para la población de bovinos de carne estudiada en este trabajo. Por otro lado, para PN las h_d^2 reportadas en la literatura se encuentran comprendidas entre 0 y 0,39 y para PD se reportan valores similares, que oscilan entre 0 a 0,56 (Rodriguez-Almeida *et al.*, 1995; Gengler, 1997). Nótese que los valores de h_d^2 obtenidos para PN y PD en este capítulo se encuentran comprendidos dentro de este rango de valores. Bolormaa *et al.* (2015) reportaron una h_d^2 de 0,11 para PD empleando información genómica, valor inferior a algunos obtenidos empleando información genealógica únicamente y superior al obtenido en este trabajo.

Por otro lado, numerosos estudios llevados a cabo en poblaciones de bovinos de carne demostraron que la consanguinidad afecta negativamente a los caracteres de crecimiento. La depresión consanguínea ha sido estimada en reiteradas oportunidades para diferentes poblaciones. De hecho, las estimaciones obtenidas para los tres caracteres de crecimiento en la población de bovinos de carne estudiada son comparables en orden de magnitud a valores reportados en la literatura. Por ejemplo para el caso de PD se han reportados valores para la depresión consanguínea que oscilan entre -0,44 kg y -0,896 kg por cada 1 % de aumento en la consanguinidad (Burrow, 1993 y 1998; Santana *et al.*, 2010; Falcão *et al.*, 2001). Nótese que los resultados presentados en esta tesis para dicho carácter son escasamente superiores a estos alcanzando valores de aproximadamente −1 kg por cada 1% de aumento de la consanguinidad. Por otro lado, los resultados reportados en la literatura para PN en poblaciones de bovinos de carne son más variables. Burrow (1998) reportó que dicho carácter no se ve afectado por la consanguinidad, mientras que otros obtuvieron estimaciones comprendidas entre -0,074 kg y -0,38 kg por cada 1% de aumento de la consanguinidad (Burrow, 1993; Swiger *et al.*, 1961). Los resultados obtenidos en este trabajo para la población de bovinos de carne estudiada se encuentran comprendidos entre

todos los valores reportados, siendo -0,048 kg por cada 1% de aumento en la consanguinidad.

Capítulo 6

Discusión general

Discusión general

La presente tesis se centró, por un lado, en el empleo de los MF en modelos de evaluación genómica ssGBLUP. Las contribuciones de esta tesis, en relación a dicha temática, fueron de índole teórica, metodológica y de aplicación. Por un lado se mostró la relación teórica entre el parámetro de relación ancestral (γ) con las (co)varianzas de las frecuencias alélicas de la población base y, en consecuencia, con el índice de fijación F_{st} (Wright, 1943). Además, se estableció la relación entre el parámetro s y el número de marcadores empleados en el análisis, permitiendo contar con un modo sencillo y rápido de calcularlo. Dichos resultados teóricos dieron lugar a las contribuciones metodológicas y de aplicación para la estimación de los parámetros s y γ. Para este último se propusieron y evaluaron tres métodos de estimación. El objetivo fue hallar metodologías que permitieran un procedimiento de estimación exacto y computacionalmente eficiente. Los resultados obtenidos permitieron generar recomendaciones para el usuario al momento de emplear la metodología y estimar los parámetros necesarios. Además, dentro de los aportes de aplicación, se evaluó el empleo de MF en los modelos ssGBLUP en términos predictivos y se mostró por simulación estocástica que esta metodología permite mantener los niveles de exactitud y mejorar significativamente en términos de sesgo comparado con el ssGBLUP tradicional.

Además se contribuyó a la literatura sobre modelos de evaluación genómica que incluyen dominancia. Aquí el aporte principal consistió en evaluar la posibilidad de incluir los efectos génicos de dominancia dentro de los modelos que incorporan la información genómica para caracteres de crecimiento de bovinos de carne. Existen escasos trabajos que aborden esta temática en hacienda de carne. Se estimaron las varianzas de dominancia para tres caracteres de crecimiento como así también la depresión consanguínea para cada uno. A continuación se sintetizan los resultados obtenidos, destacando aquellos de particular interés para la disciplina. Para una discusión más organizada, se describen dichas contribuciones siguiendo el orden en que fueron presentadas en las diferentes secciones de la tesis.

La principal motivación de esta investigación fue el intenso empleo de la información genómica en la mejora genética animal durante la última década, hecho que introdujo un cambio paradigmático en la predicción del mérito genético en la disciplina. Hoy en día la información genómica consiste en el genotipado con paneles de marcadores SNP de alta densidad. El desafío de los mejoradores es buscar alternativas para incorporar dicha información a las evaluaciones genéticas de modo eficiente con el objetivo de incrementar la exactitud de predicción. A tal fin, se han propuesto varias metodologías que permitieron alcanzar avances notorios (García-Ruiz *et al.*, 2016) si bien existen argumentos

teóricos para discutir cómo se utilizan conjuntamente la información del pedigrí y la genómica en métodos cómo GBLUP y ssGBLUP (Thompson, 2013), hecho que requiere evitar la redundancia (multicolinealidad) entre ambas fuentes para obtener un predictor con propiedades de mínima varianza (Cantet *et al.*, 2017). En cuanto hace a recuperar la información perdida de posibles parentescos entre fundadores, Legarra *et al.* (2015) sugirieron emplear MF, metodología que hemos discutido en profundidad en capítulos anteriores. Cabe destacar que el constante crecimiento en volumen de los datos genómicos disponibles requiere soluciones eficientes y sencillas a la hora de implementarlas. En términos generarles, los aportes más relevantes de esta tesis se dieron en este aspecto, evaluando y proponiendo alternativas que permitan optimizar el uso de la información disponible con el objetivo de sacar el mayor provecho posible de los datos para mejorar la calidad de las predicciones en las evaluaciones genómicas.

Las principales contribuciones del capítulo 2 son de naturaleza teórica y metodológica. En él se presentó la teoría de Legarra *et al.* (2015) como una aplicación desarrollada sobre la base de los trabajos de Jacquard (1969, 1974) VanRaden (1992), Aguilar y Misztal (2008), VanRaden *et al.* (2011), Colleau y Sargolzaei (2011) y Christensen (2012). Empleando este marco teórico, los parámetros centrales del modelo con MF son γ y s. Entre los principales resultados se destaca la relación entre el parámetro γ y las covarianzas de las frecuencias alélicas base y, consecuentemente, la relación con el índice de fijación F_{st}. La importancia de este resultado radica en dos aspectos. El primero está dado por la relevancia del parámetro γ por sí mismo, más allá de su utilidad en términos predictivos. Dada su relación con F_{st}, γ es un valor característico de la variabilidad de una población en particular. En adición, el resultado posee importancia en términos prácticos respecto a la implementación y adopción del método. Esto se debe a que la relación con las covarianzas de las frecuencias alélicas base permite una sencilla estimación de γ, facilitando así la implementación de los MF en las evaluaciones genómicas.

De hecho, una vez establecida esta relación entre el parámetro γ y las covarianzas de las frecuencias alélicas base, en el capítulo 3 se propusieron y evaluaron mediante simulación métodos alternativos para estimar γ. Los métodos propuestos contemplan varios escenarios factibles de observar en casos reales, en los que pueden contarse con una única población o varias. De los cuatro métodos de estimación de γ evaluados, entre los que se incluye el propuesto inicialmente por Legarra *et al.* (2015), tanto ML como GLS presentaron el mejor desempeño y, consecuentemente, son los que se recomienda emplear.

Luego, en el capítulo 4 se evaluó la factibilidad de incorporar un MF en el modelo ssGBLUP y se comparó su desempeño en términos predictivos con el de ssGBLUP tradicional, ssGBLUP incorporando la consanguinidad al construir $\boldsymbol{A}^{-1}$ y BLUP clásico sin considerar información genómica. El modelo con el MF presentó un mejor desempeño en términos de sesgo, observándose una disminución significativa del mismo, manteniendo los niveles de exactitud comparado a las otras variantes de ssGBLUP. Estos resultados permiten afirmar que es factible implementar el uso de los MF en la evaluación genómica

con beneficios dados por la mejora en la compatibilidad de las matrices $\boldsymbol{A}$ y $\boldsymbol{G}$ y, consecuentemente, con un impacto positivo en el sesgo de predicción. Recordemos que minimizar el sesgo puede tener impactos importantes a la hora de llevar a cabo la selección. Según el esquema de selección, las predicciones del mérito genético sesgadas pueden impulsar decisiones de selección subóptimas con impacto directo en la respuesta a la selección. Por ejemplo, en una evaluación genética sesgada, Henderson (1973) mostró que los padres de generaciones recientes se encontraban sub-evaluados con respecto a aquellos más viejos. Por otro lado, es importante destacar que la implementación de ssGBLUP con MF es sencilla. En tal sentido, algunos resultados teóricos y metodológicos de esta tesis permiten calcular γ de modo sencillo, con escaso esfuerzo computacional y asegurando una alta exactitud de estimación. Ambos aspectos son altamente deseables a la hora de proponer una implementación sencilla del método en términos predictivos.

Otro aspecto a considerar es el alcance de los resultados obtenidos. Los resultados teóricos y metodológicos presentados en el capítulo 2 y 3 son generales y se pueden aplicar tanto a casos con una población con una única base o, en su defecto a aquellas con más de una, buscando abarcar una amplia gama de situaciones posibles. Ahora bien, al momento de evaluar la performance de los métodos de estimación de γ, así como también el desempeño en términos predictivos del modelo con MF se empleó una única población simulada. Dicha simulación permitió determinar el mejor método de estimación paramétrica y generar recomendaciones generales para decidir cuál metodología emplear. Cabe destacar que en esta tesis se abordó la primera etapa de evaluación del modelo con MF, motivo por el cual la estructura poblacional empleada en la simulación fue relativamente sencilla, pero representativa de casos reales frecuentes. El objetivo fue verificar la viabilidad del método y sus beneficios potenciales con el objetivo de sentar una base sobre la cual continuar investigando. En tal sentido, una vez obtenidos los resultados presentados en el capítulo 3 y 4, y sobre la base de los mismos, trabajos posteriores como los de Xiang *et al.* (2017), Vandenplas *et al.* (2017) y Van Grevenhof *et al.* (2017) evaluaron el desempeño del método en otros escenarios más complejos, como ser con datos reales, poblaciones con más de una base y con animales cruza. Estos casos permitieron cubrir un mayor número de escenarios posibles en los que puede emplearse un modelo con MF y evaluar su desempeño bajo diferentes condiciones. En dichos trabajos se utilizaron las metodologías propuestas y evaluadas en esta tesis para estimar γ. En términos predictivos, los resultados reportados por estos autores respaldan a aquellos presentados en este trabajo con datos simulados. Ahora bien, lejos de agotarse el tema, todavía existen varios escenarios reales, de complejidad creciente, en los que los MF podrían contribuir significativamente a la hora de llevar a cabo evaluaciones genómicas. Consecuentemente, sería de gran interés en el futuro estudiar más casos reales, sean escenarios simples como poblaciones con una única base, o complejos como razas puras y sus cruzas (Christensen *et al.*, 2014; Lourenco *et al.*, 2016) y poblaciones selectas con UPG (Quaas, 1988) para evaluar el empleo de los MF y su impacto en la calidad de las predicciones.

En relación con la dominancia, los resultados obtenidos al incorporar este efecto génico en modelos de evaluación genómica para caracteres de crecimiento en bovinos de carne indicaron que su inclusión no generó ningún beneficio significativo. Esto se observó tanto en términos de bondad de ajuste de los modelos, como así también en relación a las estimaciones de las varianzas de dominancia obtenidas para los tres caracteres de crecimiento analizados. En todos los casos, estos últimos no superaron el 10% del valor de la varianza aditiva, lo que sugiere escasa variación genética no aditiva debida a la dominancia para los caracteres de crecimiento en la población de bovinos de carne estudiada. Los resultados obtenidos son comparables a algunos de los reportados en la literatura, tanto para los componentes de varianza como para la depresión consanguínea de cada carácter.

Recientemente se ha renovado el interés en la posibilidad de considerar efectos no aditivos dentro de los modelos de evaluación genómica y, en consecuencia, existe en la actualidad una activa investigación en el tema en diferentes especies de animales domésticos. Sin embargo, para el caso puntual de ganado bovino de carne, es muy escasa la investigación que se ha generado referida al tema. Esto puede atribuirse principalmente a la falta de grandes bases de datos con un gran número de animales con información genómica y fenotípica, tal como discute Varona *et al.* (2018). Es en este aspecto que radica la principal contribución el capítulo 5. En el mismo se empleó una gran base de datos de una población real para obtener estimaciones de las varianzas de dominancia y determinar si es conveniente incorporar la dominancia al modelo. Los resultados obtenidos se suman a aquellos reportados por Bolormaa *et al.* (2015) para bovinos de carne empleando información genómica. Es importante destacar que en éste último trabajo se presentaron estimaciones para varios caracteres, pero sólo dos de ellos involucraba crecimiento. Por el contrario, esta tesis se enfocó en este tipo de caracteres dado que determinan, en gran medida, la productividad e ingresos económicos de un sistema productivo de hacienda de carne. Además, Bolormaa *et al.* (2015) llevaron a cabo sus análisis empleando información de animales de diferentes razas y sus cruzas, mientras que en esta tesis se empleó una extensa base de datos genómica y fenotípica de individuos de la misma raza pura. Consecuentemente, se puede afirmar que el análisis y los resultados reportados en nuestro trabajo son complementarios a los presentados por Bolormaa *et al.* (2015).

Cabe mencionar que, si bien la base de datos de la población de bovinos de carne empleada contaba con un gran número de registros fenotípicos y de animales genotipados, no fue lo suficientemente informativa como para evitar un confundimiento con los efectos maternos de los caracteres. En principio sería deseable para futuras investigaciones, contar con un mayor número de registros con varias crías por madres y que además cuenten con fenotipo e información genómica. Además, contar con una proporción mayor de relaciones de hermanos enteros podría contribuir de modo considerable a la estimación dado que este tipo de parentesco es el más informativo en términos de relación de dominancia. Por otro lado, podrían incluirse otros caracteres de interés en el análisis, como por ejemplo aquellos asociados con la fertilidad cuyo impacto económico y productivo es también considerable

en el sistema de producción de carne. Además, dada la naturaleza de dichos caracteres (relacionados a la adaptabilidad, del inglés "*fitness traits*"), es esperable una mayor influencia de la dominancia y, consecuentemente, mayores beneficios potenciales de considerarla dentro del modelo de evaluación genómica.

Capítulo 7

Conclusiones

Conclusiones

Sobre la base de toda la investigación realizada, las principales conclusiones de esta tesis se sintetizan a continuación:

1. Se presentaron contribuciones teóricas y metodológicas para implementar un modelo de selección genómica incorporando MF. Se mostró que los parámetros de relaciones ancestrales (γ) son proporcionales a las (co)varianzas estandarizadas de las frecuencias alélicas base entre poblaciones y su relación con el índice de fijación F_{ST} (Wright, 1943). Se mostró también que el parámetro *s* puede calcularse simplemente como la mitad del número de marcadores empleados en el análisis. Dichos aportes teóricos contribuyen a generar una implementación sencilla del método, priorizando la facilidad de cálculo de los parámetros centrales, lo que permite facilitar la difusión y adopción del método.

2. Se propusieron y evaluaron metodologías para calcular el parámetro γ en función de las (co)varianzas estimadas de las frecuencias alélicas de la base con el objetivo de simplificar el uso del modelo con MF. Estas metodologías surgieron como alternativas a las inicialmente propuestas por Christensen (2012) y Legarra *et al.* (2015) y contemplan la posibilidad de trabajar con una o varias poblaciones base. Se verificó por medio de una simulación de una población bajo selección que, tanto GLS como ML son insesgados y computacionalmente eficientes. Estos resultados permiten recomendar el empleo de cualquiera de ambos métodos de estimación de los γ para una eventual implementación del modelo en un programa de selección genómica.
3. Se verificó por simulación que incorporar un MF en ssGBLUP permite mantener los niveles de exactitud de ssGBLUP tradicional y obtener predicciones menos sesgadas que sin la incorporación de dichos *pseudo*-individuos. Además, el modelo con MF permite obtener estimaciones de los parámetros genéticos más precisas.

4. Se evaluó la conveniencia de incorporar la dominancia en el modelo de selección genómica para caracteres de crecimiento en una población de ganado de bovinos de carne, obteniéndose estimaciones de la varianza de dominancia para tres caracteres de crecimiento. Los valores obtenidos fueron bajos, lo que permite concluir que la variación genética debida a las desviaciones de dominancia es muy reducida para estos caracteres en la población de ganado de carne estudiada. Se observó también que el ajuste del modelo no mejoró al incorporar el efecto de dominancia para ninguno de los tres caracteres. En consecuencia, los resultados muestran que para estos casos no se obtiene ningún beneficio al considerar la dominancia en el

modelo, de modo que no se recomienda hacerlo. Cabe destacar que, tal como se esperaba según los fundamentos teóricos del modelo presentados en el capítulo 5, las estimaciones de la varianza aditiva no se vieron alteradas por la inclusión de los efectos no aditivos en el modelo. Se obtuvieron también valores de la depresión consanguínea para cada uno de los caracteres comparables en orden de magnitud a aquellos reportados en la literatura para otras poblaciones de bovinos de carne.

BIBLIOGRAFÍA

Aguilar, I. y Misztal, I. 2008. Technical note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. J. Dairy Sci., 91: 1669–1672.

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S. y Lawlor, T. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci., 93: 743–752.

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G. y Hayes, B. J. 2016. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet. Sel. Evol., 48:186.

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E.y Hayes, B. J. 2017. Including non-additive genetic effects in mating programs to maximize dairy farm profitability. J. Dairy Sci., 100: 1203 – 1222.

Bolormaa, S., Pryce, J. E., Zhang, Y., Reverter, A., Barendse, W., Hayes, B. J. y Goddard, M. E. 2015. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. Genet. Sel. Evol., 47(1): 26.

Burrow, H. M. 1993. The effects of inbreeding in beef cattle. Anim. Breed. Abstr., 61(11): 737 – 751).

Burrow, H. M. 1998. The effects of inbreeding on productive and adaptive traits and temperament of tropical beef cattle. Livest. Prod. Sci., 55(3): 227 – 243.

Cantet, R. J. C., García-Baccino, C. A., Rogberg-Muñoz, A., Forneris, N. S. y Munilla, S. 2017. Beyond genomic selection: The animal model strikes back (one generation)!. J.Anim. Breed.Genet., 134(3): 224 – 231.

Christensen, O. F. y Lund, M. S. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol., 42:2.

Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol., 44:37.

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T. y Su, G. 2012. Single-step methods for genomic evaluation in pigs. Anim. Int. J. Anim. Biosci.: 1.

Christensen, O.F., Legarra, A., Lund, M. S. y Su, G. 2015. Genetic evaluation for threeway crossbreeding. Genet Sel Evol.,47:98.

Christensen, O. F., Madsen, P., Nielsen, B. y Su, G. 2014. Genomic evaluation of both purebred and crossbred performances. Genet. Sel. Evol., 46: 1 – 9.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution, 23: 72 – 84.

Colleau, J. J. y Sargolzaei, M. 2011. MIM: an indirect method to assess inbreeding and coancestry in large incomplete pedigrees of selected dairy cattle. J. Anim. Breed. Genet., 128(3): 163 – 173.

Crow, J. y Kimura, M. 1970. An introduction to population genetics theory. Harper and Row, New York.

de Boer, I. y Hoeschele, I. 1993. Genetic evaluation methods for populations with dominance and inbreeding. Theor. Appl. Genet., 86: 245 – 258.

de Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. y Cotes, J. M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1): 375 – 385.

de los Campos, G., Sorensen, D. y Gianola, D. 2015. Genomic heritability: what is it?. PLoS Genet., 11(5): e1005048.

Duenk, P., Calus, M. P. L., Wientjes, Y. C. J. y Bijma, P. 2017. Benefits of dominance over additive models for the estimation of average effects in the presence of dominance. G3: Genes|Genomes|Genetics, 7: 3405 – 3414.

Emik, L.O. y Terrill, C.E. 1949. Systematic procedures for calculating inbreeding coefficients. J Hered., 40:51 – 5.

Ertl, J., Legarra, A., Vitezica, Z. G., Varona, L., Edel, C., Emmerling, R. y Gotz, K-U. 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol., 46:40.

Esfandyari, H., Bijma, P., Henryon, M., Christensen, O. F. y Sorensen, A. C. 2016. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet. Sel. Evol., 48:40.

Falcão, A.J.S., Filho, R.M., Magnabosco, C.U., Bozzi, R. y Lima, F.A.M. 2001. Effects of inbreeding on reproductive and growth traits, and breeding values in a closed Brown Swiss herd. Braz. J. Anim. Sci., 30: 83 – 92.

Falconer, D. S. y Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. Longman New York.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. y Servin, B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics, 193: 929 – 941.

Fernández, E.N., Legarra, A., Martínez, R., Sánchez, J.P. y Baselga, M., 2017. Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler equivalent model. J. Anim. Breed. Genet., 134: 184 –195.

Fernando, R. L., Dekkers, J.C. y Garrick, D.J. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol., 46:50.

Forneris, N. S., Legarra, A., Vitezica, Z. G., Tsuruta, S., Aguilar, I., Misztal, I. y Cantet, R. J. 2015. Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. Genetics, 199: 675 – 681.

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J. y Van Tassell, C. P. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. PNAS., 113(28): E3995 – E4004.

Gengler, N., Van Vleck, L. D., MacNeil, M. D., Misztal, I. y Pariacote, F. A. 1997. Influence of dominance relationships on the estimation of dominance variance with sire-dam subclass effects. J. Anim. Sci., 75(11): 2885 – 2891.

Gengler, N., Mayeres, P. y Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal, 1: 21 – 28.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. y Fernando, R. 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183: 347 – 363.

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A. et al. 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science, 324: 528 – 532.

Harris, B. L. y Johnson, D. L. 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci., 93: 1243 – 1252.

Heidaritabar, M., Wolc, A., Arango, J., Zeng, J., Settar, P., Fulton, J. E., et al. 2016. Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. J. Anim. Breed. Genet., 133: 334 – 346.

Henderson, C. R., 1973 Sire evaluations and genetic trends. J Anim Sci Symposium.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics., 32: 69 – 83.

Henderson, C.R. 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, Ontario, Canada.

Hill, W. G., Goddard, M. E. y Visscher, P. M. 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet., 4:e1000008.

Hill, W. G. 2010. Understanding and using quantitative genetic variation. Philos. Trans. R. Soc. Lond. B Sci., 365: 73 – 85.

Hill, W. G. 2017. “Conversion” of epistatic into additive genetic variance in finite populations and possible impact on long-term selection response. J Anim Breed Genet., 134:196 – 201.

Huang, W. y Mackay, T. F. C. 2016. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. PLoS Genet., 10:e1006421.

Jacquard, A. 1969. Evolution of genetic structures of small populations. Biodemography and social biology, 16: 143–157.

Jacquard, A., 1974 The Genetic Structure of Populations. Springer Verlag, Berlin/Heidelberg/New York.

Jiang, J., Shen, B., O', Connell, J. R., VanRaden, P. M., Cole, J. B. y Ma, L. 2017. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. BMC Genomics, 18: 425.

Kennedy, B. 1991. CR Henderson: The unfinished legacy. J. Dairy Sci. 74: 4067 – 4081.

Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., et al. 2009. A genome-wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS ONE, 4: e4668.

Legarra, A, Aguilar, I y Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci., 92: 4656 – 63.

Legarra, A., Christensen, O. F., Aguilar, I. y Misztal, I. 2014. Single step, a general approach for genomic selection. Livest. Sci., 166: 54 – 65.

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I. y Misztal, I. 2015 Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics, 200: 455 – 68.

Legarra, A. 2016. Comparing estimates of genetic variance across different relationship models. Theor Popul Biol., 107: 26 – 30.

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Chen, C. Y., Herring, W. O. y Misztal, I. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J. Anim. Sci., 94: 909.

Lynch, M. y B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Suderland, MA, USA.

Makgahlela, M. L., Strandén, I., Nielsen, U. S., Sillanpää, M. J. y Mäntysaari, E.A. 2014. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. J Dairy Sci., 97: 1117 – 27.

Maki-Tanila, A. 2007. An overview on quantitative and genomic tools for utilising dominance genetic variation in improving animal production. Agric. Food Sci. 16: 188 – 198.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie, Paris.

Mäntysaari, E. y Van Vleck, L. 1989. Restricted maximum likelihood estimates of variance components from multitrait sire models with large number of fixed effects. J. Anim. Breed. Genet., 106: 409 – 422.

Mäntysaari, E., Liu, Z. y VanRaden, P., 2010 Interbull validation test for genomic evaluations. Interbull Bull 41:17.

Masuda, Y., Misztal, I., Tsuruta, S., Legarra, A., Aguilar, I., Lourenco, D. A. L., Fragomeni, B. O. y Lawlor, T. J. 2016. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J. Dairy Sci., 99: 1968 – 1974.

McPeek, M. S., Wu, X. y Ober, C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics., 60: 359 – 67.

Mehrabani-Yeganeh, H., Gibson, J. P. y Schaeffer L. R. 2000. Including coefficients of inbreeding in BLUP evaluation and its effect on response to selection. J. Anim. Breed. Genet., 117: 145 – 151.

Meuwissen, T. y Luo, Z. 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol., 24: 305 – 313.

Meuwissen, T., Hayes, B. J. y Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819 – 1829.

Meuwissen, T., Luan, T. y Woolliams. J. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet., 128: 429 – 439.

Misztal, I., Varona, L., Culbertson, M., Gengler, N., Bertrand, J. K., Mabry, J., et al. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol. Agron. Soc. Environ., 2: 227 – 233.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. y Lee, D. H. 2002. BLUPF90 and related programs (BGF90). CD-ROM, Communication No. 28−07, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Misztal, I., Legarra, A. y Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci., 92(9): 4648 – 4655.

Misztal, I., Vitezica. Z. G., Legarra, A., Aguilar, I. y Swan, A. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet., 130: 252 – 258.

Moghaddar, N. y van der Werf, J. H. J. 2017. Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. J. Anim. Breed. Genet., 134: 453 – 462.

Nguyen, T. N. y Nagyné-Kiszlinger, H. 2016. Dominance effects in domestic populations. Acta Agraria Kaposvariensis, 20: 1 – 20.

Powell, J. E., Visscher, P. M. y Goddard, M. E. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet., 11: 800 – 805.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics, 949 – 953.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci.,71: 1338 – 45.

Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res (Camb)., 67: 175 – 85.

Robertson A. 1975. Gene Frequency Distributions as a Test of Selective Neutrality. Genetics, 81: 775 – 785.

Rodriguez-Almeida, F. A., Van Vleck, L. D., Willham, R. L. y Northcutt, S. L. 1995. Estimation of non-additive genetic variances in three synthetic lines of beef cattle using an animal model. J. Anim. Ssci., 73(4): 1002 – 1011.

Santana, Jr. M. L., Oliveira, P. S., Pedrosa, V. B., Eler, J. P., Groeneveld, E. y Ferraz, J. B. S. 2010. Effect of inbreeding on growth and reproductive traits of Nellore cattle in Brazil. Livest. Sci., 131(2–3): 212 – 217.

Sargolzaei, M. y Schenkel, F. S. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25: 680 – 681.

Sargolzaei, M., Chesnais, J. y Schenkel, F. S. 2012. Assessing the bias in top GPA bulls. 2012. cgil.uoguelph.ca/dcbgc/Agenda1209/DCBGC1209_Bias_Mehdi.pdf. Visitado el 21 de julio de 2016.

Searle, S. R., 1982. Matrix Algebra Useful for Statistics. John Wiley & Sons, Inc., NY, USA.

Silió, L., Rodríguez, M. C., Fernández, A., Barragán, C., Benítez, R., Óvilo, C. y Fernández, A. I. 2013. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J. Anim. Breed. Genet., 130(5): 349 – 360.

Spelman, R. J., Arias, J., Keehan, M. D., Obolonkin, V., Winkelman, A. M., Johnson, D. L. y Harris, B. L. 2010. Application of genomic selection in the New Zealand dairy cattle industry. En: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1 – 6 August 2010; Leipzig,

Strandén, I. y Christensen, O. F. 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43: 25.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M. y Lund, M. S. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS ONE, 7:e45293.

Sun, C., VanRaden, P.M., O'Connell, J.R., Weigel, K.A. y Gianola, D., 2013. Mating programs including genomic relationships and dominance effects. J. Dairy Sci. 96: 8014 – 8023.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A. y Meuwissen, T. H. E. 2008. Genomic selection using different marker types and densities. J. Anim. Sci., 86:2447 – 54.

Swiger, L.A., Gregory, K.E., Koch, R.M. y Arthaud, V.A. 1961. Effect of inbreeding on performance traits of beef cattle. J. Anim. Sci. 20: 626 – 630

Thompson R. 1979. Sire evaluation. Biometrics, 35:339 – 53.

Thompson E. A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics. 194: 301 – 326.

Toro, M. A. y Varona, L. 2010. A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol., 42: 33.

Toro, M. A., García-Cortés, L. A. y Legarra, A., 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet. Sel. Evol., 43: 27.

Tsuruta, S., Misztal, I., Aguilar, I. y Lawlor, T. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94: 4198 – 4204.

Vandenplas, J., Calus, M.P.L., Brinker, T., Ellen, E.D., Bink, M.C.A.M. y Ten Napel, J. 2017. Single-step GBLUP using metafounders to predict crossbred performance of

laying hens. En: 8th Annual Meeting of the European Federation of Animal Science. Tallin, Estonia.

Van Grevenhof, E.M., Vandenplas, J. y Calus., M.P.L.. 2017. Using metafounders to model purebred relationships in genomic prediction for crossbreeding. En: 8th Annual Meeting of the European Federation of Animal Science. Tallin, Estonia.

VanRaden, P., 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. J. Dairy Sci., 75: 3136 – 3144.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91: 4414 – 23.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. y Schenkel. F. S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci., 92: 16 – 24.

VanRaden, P., Olson, K., Wiggans, G., Cole, J. y Tooker, M. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94: 5673 – 5682.

Varona, L., Legarra, A., Toro, M. A. y Vitezica, Z. G. 2018. Non-additive Effects in Genomic Selection. Frontiers in genetics, 9: 78.

Visscher, P. M. 2006. A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res Hum. Genet., 9: 490 – 495.

Vitezica, Z., Aguilar, I., Misztal, I. y Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb)., 93: 357 – 66.

Vitezica, Z. G., Varona, L. y Legarra, A. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics, 195: 1223 – 1230.

Vitezica, Z. G., Legarra, A., Toro, M. A. y Varona, L. 2017. Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics, 206(3): 1297 – 1307.

Winkelman, A. M., Johnson, D. L. y Harris, B. L. 2015. Application of genomic evaluation to dairy cattle in New Zealand. J. Dairy Sci. 98: 659 – 675.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat., 56: 330 – 338.

Wright S. 1931. Evolution in Mendelian populations. Genetics, 16: 97 – 159.

Wright S. 1943. Isolation by distance. Genetics, 28: 114 – 38.

Xiang, T., Christensen, O. F., Vitezica, Z. G. y Legarra, A. 2016. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol., 48: 92.

Xiang, T., Christensen, O. F. y Legarra, A. 2017. Genomic evaluation for crossbred performance in a single-step approach with metafounders 1. J.Anim. Sci., 95(4): 1472 – 1480.

Zeng, J., Toosi, A., Fernando, R. L., Dekkers, J. C. M. y Garrick, D. J. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet. Sel. Evol., 45: 11.

APÉNDICE

El apéndice contiene detalles y desarrollos algebraicos que no se detallaron en el texto principal de la tesis.

Equivalencia de las matrices de relaciones genómicas

La matriz $\boldsymbol{G}$ (n x n) descripta por Christensen (2012) puede computarse según la siguiente expresión

$$\boldsymbol{G} = 2\left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' \Big/ n \qquad \text{[A.1]}$$

donde $\boldsymbol{M}$ (n x m) contiene genotipos codificados como {0,1,2} y $\boldsymbol{J}$ (n x m) es una matriz de elementos iguales a uno. El objetivo de este párrafo es mostrar la relación lineal que existe entre esta matriz con aquella que describe coeficientes IBS. De hecho,

$$\boldsymbol{G}_{\text{IBS}} = \boldsymbol{G}/2 + \boldsymbol{11}' \qquad \text{[A.2]}$$

$\boldsymbol{G}_{\text{IBS}}$ (n x n) puede describirse en términos de identidad o conteos (Ritland, 1996; Toro *et al*., 2011):

$$\boldsymbol{G}_{\text{IBS}_{ij}} = \frac{1}{n}\sum_{m=1}^{n} 2\frac{\sum_{k=1}^{2}\sum_{l=1}^{2} I_{kl}}{4} \qquad \text{[A.3]}$$

donde I_{kl} mide la identidad (puede tomar valores 1 o 0) del alelo k en el individuo i con el alelo l en el individuo j, y las medidas de identidad de un sólo locus son promediadas a través de los k loci. Existe una expresión algebraica para este "conteo". Toro et al (2011) en su expresión (1) muestran que, para marcadores bialélicos, para un locus k (se omite en l expresión para mayor claridad):

$$f_{M_{\text{ij}}} = \frac{m_i}{2}\frac{m_j}{2} + \left(1 - \frac{m_i}{2}\right)\left(1 - \frac{m_j}{2}\right), \qquad \text{[A.4]}$$

para la coancestría (un medio de la relación) $f_{M_{\text{ij}}}$ de los individuos i y j , donde $m/2$ es la "frecuencia génica" del individuo (un medio del contenido génico (m), es decir {0,1/2,1} para los tres genotipos).

Para probar $\boldsymbol{G}_{\text{IBS}} = \boldsymbol{G}/2 + \boldsymbol{11}'$, es necesario primero traducir la ecuación en Toro et al (2011) a una escala más familiar de relaciones $g_{IBS_{ij}} = 2f_{M_{ij}}$ y contenidos génicos *m*. En consecuencia,

$$g_{IBS_{ij}} = 2f_{M_{ij}} = 2\left(\frac{m_i}{2}\frac{m_j}{2} + \left(\frac{2}{2} - \frac{m_i}{2}\right)\left(\frac{2}{2} - \frac{m_j}{2}\right)\right) \qquad [A.5]$$

$$g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2 \qquad [A.6]$$

Esta expresión puede verificarse fácilmente en un cuadro con nueve posibles genotipos:

	AA	**Aa**	**aa**
AA	2	1	0
Aa	1	1	1
aa	0	1	2

Además,

$$g_{IBS_{ij}} = m_i m_j - m_i - m_j + 2 = \left(m_i - 1\right)\left(m_j - 1\right) + 1 \qquad [A.7]$$

que se extiende a todos los individuos y al promediar entre todos los loci puede escribirse como:

$$\boldsymbol{G}_{IBS} = \frac{1}{n}\left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' + \boldsymbol{11}' \qquad [A.8]$$

En consecuencia, la matriz $\boldsymbol{G}_{IBS} = \frac{1}{n}\left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)' + \boldsymbol{11}'$ y, dado que $\boldsymbol{G} = \frac{2}{n}\left(\boldsymbol{M} - \boldsymbol{J}\right)\left(\boldsymbol{M} - \boldsymbol{J}\right)'$, se sigue que $\boldsymbol{G}_{IBS} = \frac{1}{2}\boldsymbol{G} + \boldsymbol{11}'$. La equivalencia puede verificarse también, que para los nueve genotipos. El producto cruzado $\left(m_i - 1\right)\left(m_j - 1\right)$ en el cuadro que se presenta a continuación es idéntico a $g_{IBS_{ij}} - 1$ en el cuadro presentado previamente.

	AA	**Aa**	**aa**
AA	1	0	–1
Aa	0	0	0
aa	–1	0	1